Model Selection and Goodness of Fit

G. Jogesh Babu

Penn State University
http://www.stat.psu.edu/∼babu

Director of Center for Astrostatistics

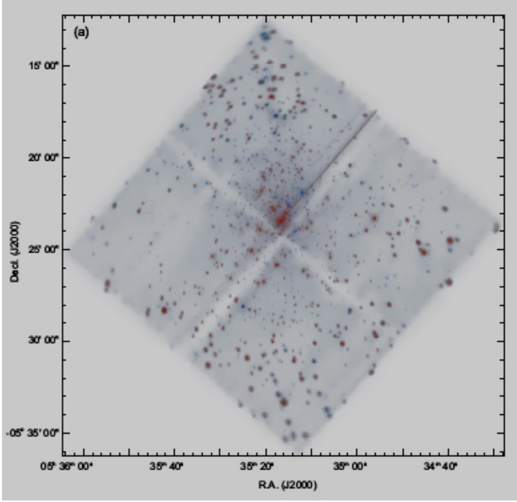
http://astrostatistics.psu.edu

Astrophysical Inference from astronomical data

Fitting astronomical data

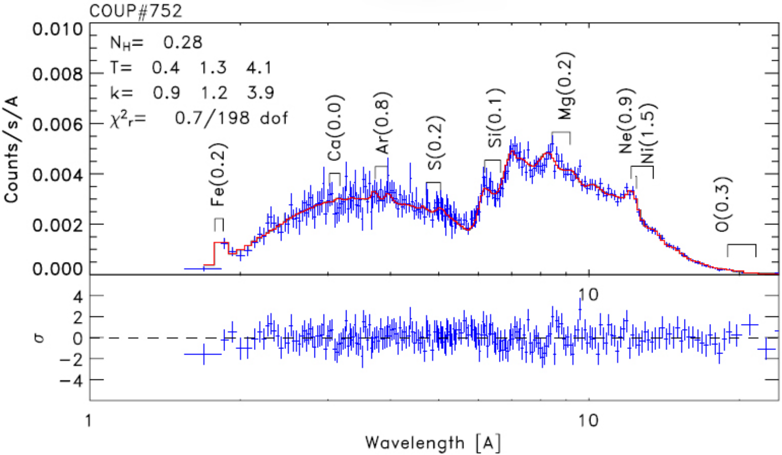
- Non-linear regression
- Density (shape) estimation
- Parametric modeling
    - Parameter estimation of assumed model
    - Model selection to evaluate different models
        - Nested (in quasar spectrum, should one add a broad absorption line BAL component to a power law continuum).
        - Non-nested (is the quasar emission process a mixture of blackbodies or a power law?).
- Goodness of fit

Chandra Orion Ultradeep Project (**COUP**)

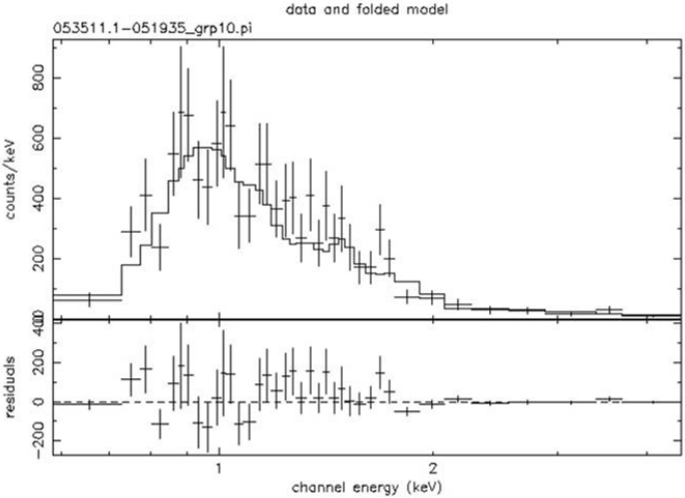


$4Bn Chandra X-Ray observatory NASA 1999
1616 Bright Sources. Two weeks of observations in 2003

What is the underlying nature of a stellar spectrum?



Successful model for high signal-to-noise X-ray spectrum.
Complicated thermal model with several temperatures
and element abundances (17 parameters)



COUP source # 410 in Orion Nebula with 468 photons
Fitting binned data using χ^2
Thermal model with absorption $A_V \sim 1$ mag

Best-fit model: A plausible emission mechanism

- Model assuming a single-temperature thermal plasma with solar abundances of elements. The model has three free parameters denoted by a vector θ.
  - plasma temperature
  - line-of-sight absorption
  - normalization
- The astrophysical model has been convolved with complicated functions representing the sensitivity of the telescope and detector.
- The model is fitted by minimizing chi-square with an iterative procedure.

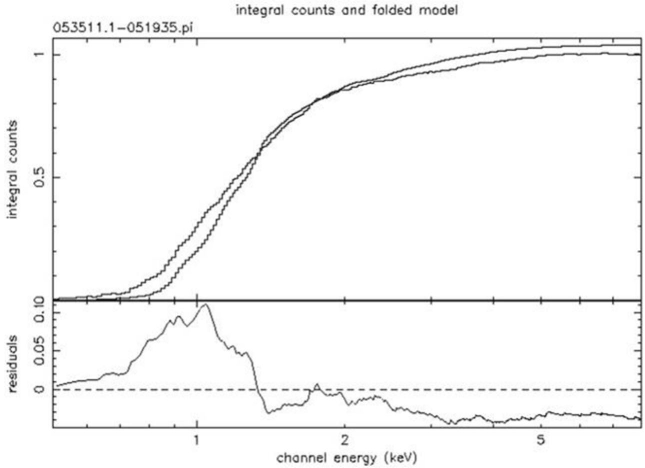
$$\hat{\theta} = \arg\min_{\theta} \chi^2(\theta) = \arg\min_{\theta} \sum_{i=1}^{N} \left(\frac{y_i - M_i(\theta)}{\sigma_i} \right)^2 .$$

Chi-square minimization is a misnomer. It is parameter estimation by *weighted least squares*.

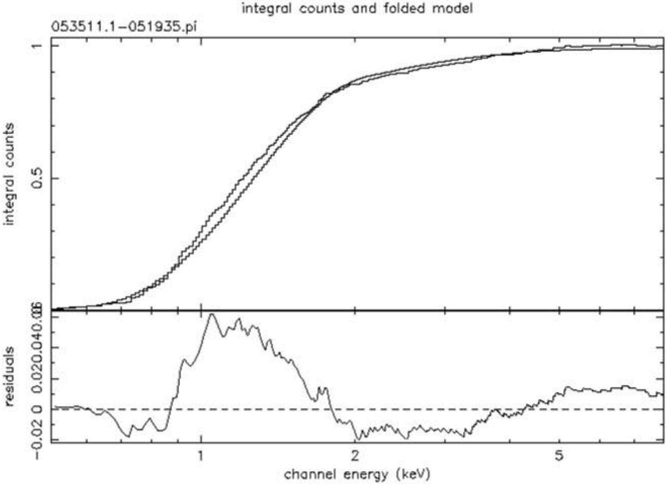
Limitations to χ^2 'minimization'

- Depends strongly on Gaussian assumptions.
- Fails when the errors are non-Gaussian (*e.g.* small-N problems with Poissonian errors).
- Does not provide clear procedures for adjudicating between models with different numbers of parameters (*e.g.* one- vs. two-temperature models) or between different acceptable models (*e.g.* local minima in $\chi^2(\theta)$ space).
- Unsuitable to obtain confidence intervals on parameters when complex correlations between the estimators of parameters are present (*e.g.* non-parabolic shape near the minimum in $\chi^2(\theta)$ space).

Alternative approach to the model fitting



Fitting to unbinned EDF
Correct model family, incorrect parameter value
Thermal model with absorption set at $A_V \sim 10$ mag



Misspecified model family!
Power law model with absorption set at $A_V \sim 1$ mag
Can the power law model be excluded with 99% confidence

Model Fitting in Astronomy to answer:

- Is the underlying nature of an X-ray stellar spectrum a non-thermal power law or a thermal gas with absorption?
- Are the fluctuations in the cosmic microwave background best fit by Big Bang models with dark energy or with quintessence?
- Are there interesting correlations among the properties of objects in any given class (e.g. the Fundamental Plane of elliptical galaxies), and what are the optimal analytical expressions of such correlations?

Model Selection in Astronomy

- Interpreting the spectrum of an accreting black hole such as a quasar. Is it a nonthermal power law, a sum of featureless blackbodies, and/or a thermal gas with atomic emission and absorption lines?

- Interpreting the radial velocity variations of a large sample of solar-like stars. This can lead to discovery of orbiting systems such as binary stars and exoplanets, giving insights into star and planet formation.

- Interpreting the spatial fluctuations in the cosmic microwave background radiation. What are the best fit combinations of baryonic, Dark Matter and Dark Energy components? Are Big Bang models with quintessence or cosmic strings excluded?

A good model should be

- Parsimonious (model simplicity)
- Conform fitted model to the data (goodness of fit)
- Easily generalizable.
- Not *under-fit* that excludes key variables or effects
- Not *over-fit* that is unnecessarily complex by including extraneous explanatory variables or effects.
- Under-fitting induces bias and over-fitting induces high variability.

A good model should balance the competing objectives of conformity to the data and parsimony.

Model Selection Framework

- Observed data D
- $M_1, \ldots, M_k$ are models for D under consideration
- Likelihood $f(D|\theta_j; M_j)$ and loglikelihood $\ell(\theta_j) = \log f(D|\theta_j; M_j)$ for model M_j.
  - $f(D|\theta_j; M_j)$ is the probability density function (in the continuous case) or probability mass function (in the discrete case) evaluated at the data D.
  - θ_i is a p_j dimensional parameter vector.

Example

$D = (X_1, \ldots, X_n)$, X_i, i.i.d. $N(\mu, \sigma^2)$ r.v. Likelihood

$$f(D|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \ \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right\}$$

Most of the methodology can be framed as a comparison between two models M_1 and M_2.

Nested Models

The model M_1 is said to be nested in M_2, if some coordinates of θ_1 are fixed, *i.e.* the parameter vector is partitioned as

- $\theta_2 = (\alpha, \gamma)$ and $\theta_1 = (\alpha, \gamma_0)$
- γ_0 is some known fixed constant vector.

Comparison of M_1 and M_2 can be viewed as a classical hypothesis testing problem of $H_0 : \gamma = \gamma_0$.

Example

M_2 Gaussian with mean μ and variance σ^2
M_1 Gaussian with mean 0 and variance σ^2

The model selection problem here can be framed in terms of statistical hypothesis testing $H_0 : \mu = 0$, with free parameter σ.

Hypothesis testing for model selection

Hypothesis testing is a criteria used for comparing two models. Classical testing methods are generally used for nested models.

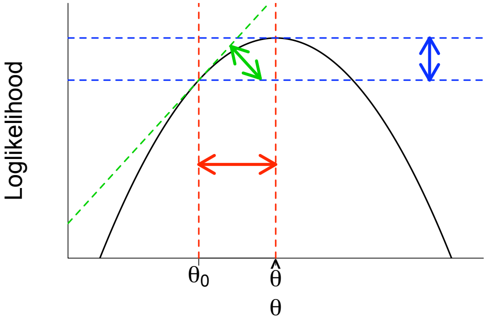
Caution/Objections

- M_1 and M_2 are not treated symmetrically as the null hypothesis is M_1.
- Cannot *accept* H_0
- Can only reject or fail to reject H_0.
- Larger samples can detect the discrepancies and more likely to lead to rejection of the null hypothesis.

The “Holy Trinity” of hypotheses tests

$H_0 : \theta = \theta_0$, $\quad\hat{\theta}$ MLE

$\ell(\theta)$ loglikelihood at θ



Wald Test

Based on the (standardized) distance between θ_0 and $\hat{\theta}$

Likelihood Ratio Test

Based on the distance from $\ell(\theta_0)$ to $\ell(\hat{\theta})$.

Rao Score Test

Based on the gradient of the loglikelihood (called the score function) at θ_0.

These three MLE based tests are equivalent to the first order of asymptotics, but differ in the second order properties.
No single test among these is uniformly better than the others.

Wald Test Statistic

$$W_n = (\hat{\theta}_n - \theta_0)^2 / Var(\hat{\theta}_n) \sim \chi^2$$

- The standardized distance between θ_0 and the MLE $\hat{\theta}_n$.
- In general $Var(\hat{\theta}_n)$ is unknown
- $Var(\hat{\theta}) \approx 1/I(\hat{\theta}_n)$, $I(\theta)$ is the Fisher's information
- Wald test rejects $H_0 : \theta = \theta_0$ when $I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)^2$ is large.

Likelihood Ratio Test Statistic

$$\ell(\hat{\theta}_n) - \ell(\theta_0)$$

Rao's Score (Lagrangian Multiplier) Test Statistic

$$S(\theta_0) = \frac{1}{nI(\theta_0)} \left(\sum_{i=1}^{n} \frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)} \right)^2$$

$X_1, \ldots, X_n$ are independent random variables with a common probability density function $f(.;\theta)$.

Example

In the case of data from normal (Gaussian) distribution

$$f(y;(\mu,\sigma^2)) = \frac{1}{\sqrt{2\pi}\sigma}\ \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

$$S(\theta_0) = \frac{1}{nI(\theta_0)}\left(\sum_{i=1}^{n}\frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)}\right)^2$$

Regression Context

$y_1,\ldots,y_n$ data with Gaussian residuals, then the loglikelihood ℓ is

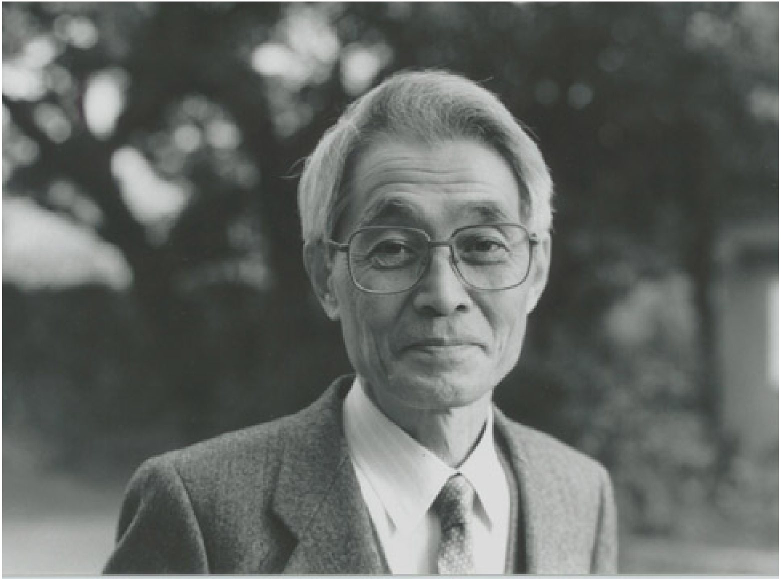
$$\ell(\beta) = \log\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\ \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mathbf{x}_i'\beta)^2\right\}$$

If M_1 is nested in M_2, then the largest likelihood achievable by M_2 will always be larger than that of M_1. Adding a a penalty on larger models would achieve a balance between over-fitting and under-fitting, leading to the so called Penalized Likelihood approach.

Information Criteria based model selection – AIC

- The traditional maximum likelihood paradigm provides a mechanism for estimating the unknown parameters of a model having a specified dimension and structure.
- Hirotugu Akaike extended this paradigm in 1973 to the case, where the model dimension is also unknown.
- Grounding in the concept of entropy, Akaike proposed an information criterion (AIC), now popularly known as Akaike's Information Criterion, where both model estimation and selection could be simultaneously accomplished.
- AIC for model M_j is $2\ell(\hat{\theta}_j) - 2k_j$. The term $2\ell(\hat{\theta}_j)$ is known as the goodness of fit term, and $2k_j$ is known as the penalty.
- The penalty term increase as the complexity of the model grows.

AIC is generally regarded as the first model selection criterion. It continues to be the most widely known and used model selection tool among practitioners.



Hirotugu Akaike (1927-2009)

Advantages of AIC

- Does not require the assumption that one of the candidate models is the "true" or "correct" model.
- All the models are treated symmetrically, unlike hypothesis testing.
- Can be used to compare nested as well as non-nested models.
- Can also be used to compare models based on different families of probability distributions.

Disadvantages of AIC

- Large data are required especially in complex modeling frameworks.
- Not *consistent*. That is, if p_0 is the correct number of parameters, and $\hat{k} = k_i \ \ (i = \arg\max_j \ 2\ell(\hat{\theta}_j) - 2k_j)$, then $\lim_{n \to \infty} P(\hat{k} > k_0) > 0$. That is even if we have very large number of observations, $\hat{p}$ does not approach the true value.

Bayesian Information Criterion (BIC)

BIC is also known as the Schwarz Bayesian Criterion

$$2\ell(\hat{\theta}_j) - k_j \log n$$

- BIC is consistent unlike AIC
- Like AIC, the models need not be nested to use BIC
- AIC penalizes free parameters less strongly than does the BIC

- Conditions under which these two criteria are mathematically justified are often ignored in practice.
- Some practitioners apply them even in situations where they should not be applied.

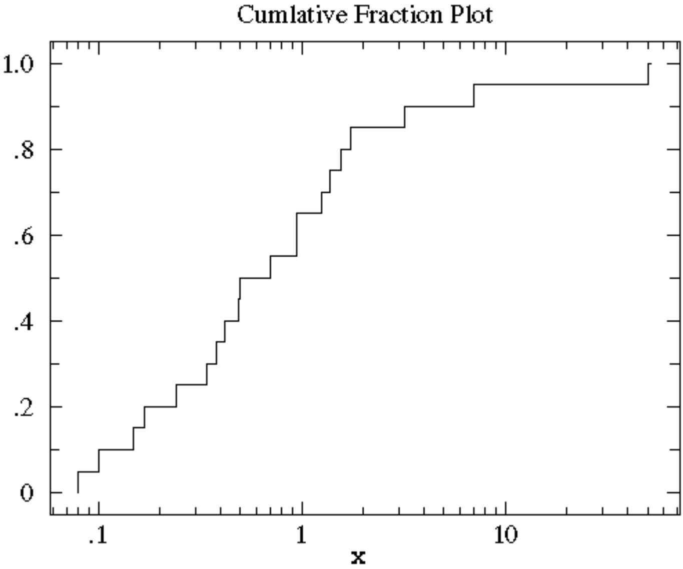
Caution

Sometimes these criteria are given a minus sign so the goal changes to finding the minimizer.

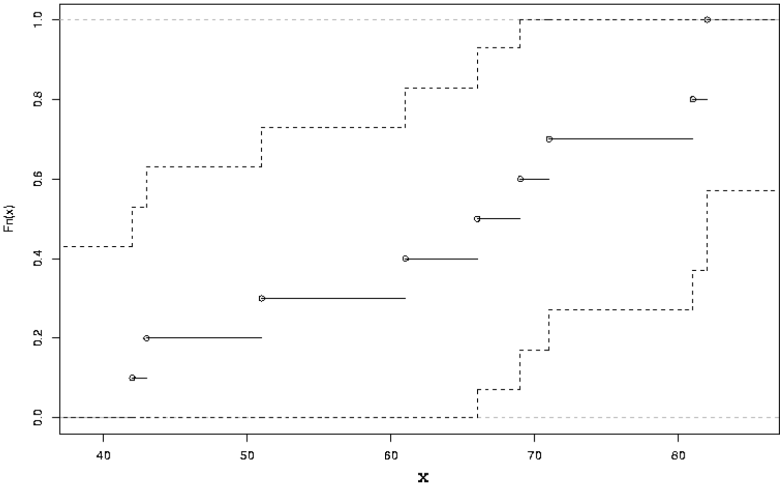
Bootstrap for Goodness of Fit

1. Astrophysical Inference from astronomical data
2. Bootstrap for Goodness of Fit
3. Statistics based on EDF
4. Processes with estimated parameters
5. Bootstrap
   - Parametric bootstrap
   - Nonparametric bootstrap
6. Confidence limits under misspecification

Empirical Distribution Function



K-S Confidence bands



$$F = F_n \pm D_n(\alpha)$$

Statistics based on EDF

Kolmogrov-Smirnov: $\sup_x |F_n(x) - F(x)|,$

$$\sup_x (F_n(x) - F(x))^+, \ \ \sup_x (F_n(x) - F(x))^-$$

Cramér-von Mises: $\int (F_n(x) - F(x))^2 \, dF(x)$

Anderson - Darling: $\int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \, dF(x)$

- These statistics are distribution free if F is continuous & univariate.
- No longer distribution free if either F is not univariate or parameters of F are estimated.

Kolmogorov-Smirnov Table

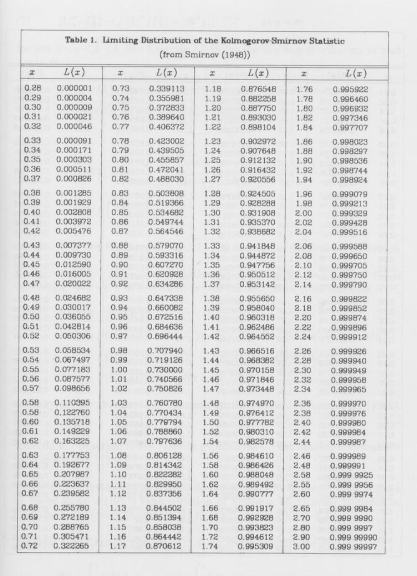


Table 1. Limiting Distribution of the Kolmogorov-Smirnov Statistic

(from Smirnov (1948))

x	$L(x)$	x	$L(x)$	x	$L(x)$	x	$L(x)$
0.28	0.000001	0.73	0.339113	1.18	0.876548	1.76	0.995922
0.29	0.000004	0.74	0.355981	1.19	0.882258	1.78	0.996460
0.30	0.000009	0.75	0.372833	1.20	0.887750	1.80	0.996932
0.31	0.000021	0.76	0.389640	1.21	0.893030	1.82	0.997346
0.32	0.000046	0.77	0.406372	1.22	0.898104	1.84	0.997707
0.33	0.000091	0.78	0.423002	1.23	0.902972	1.86	0.998023
0.34	0.000171	0.79	0.439505	1.24	0.907648	1.88	0.998297
0.35	0.000303	0.80	0.455857	1.25	0.912132	1.90	0.998536
0.36	0.000511	0.81	0.472041	1.26	0.916432	1.92	0.998744
0.37	0.000826	0.82	0.488030	1.27	0.920556	1.94	0.998924
0.38	0.001285	0.83	0.503808	1.28	0.924505	1.96	0.999079
0.39	0.001929	0.84	0.519366	1.29	0.928288	1.98	0.999213
0.40	0.002808	0.85	0.534682	1.30	0.931908	2.00	0.999329
0.41	0.003972	0.86	0.549744	1.31	0.935370	2.02	0.999428
0.42	0.005476	0.87	0.564546	1.32	0.938682	2.04	0.999516
0.43	0.007377	0.88	0.579070	1.33	0.941848	2.06	0.999588
0.44	0.009730	0.89	0.593316	1.34	0.944872	2.08	0.999650
0.45	0.012590	0.90	0.607270	1.35	0.947756	2.10	0.999705
0.46	0.016005	0.91	0.620928	1.36	0.950512	2.12	0.999750
0.47	0.020022	0.92	0.634286	1.37	0.953142	2.14	0.999790
0.48	0.024682	0.93	0.647338	1.38	0.955650	2.16	0.999822
0.49	0.030017	0.94	0.660082	1.39	0.958040	2.18	0.999852
0.50	0.036055	0.95	0.672516	1.40	0.960318	2.20	0.999874
0.51	0.042814	0.96	0.684636	1.41	0.962486	2.22	0.999896
0.52	0.050306	0.97	0.696444	1.42	0.964552	2.24	0.999912
0.53	0.058534	0.98	0.707940	1.43	0.966516	2.26	0.999926
0.54	0.067497	0.99	0.719126	1.44	0.968382	2.28	0.999940
0.55	0.077183	1.00	0.730000	1.45	0.970158	2.30	0.999949
0.56	0.087577	1.01	0.740566	1.46	0.971846	2.32	0.999958
0.57	0.098656	1.02	0.750826	1.47	0.973448	2.34	0.999965
0.58	0.110395	1.03	0.760780	1.48	0.974970	2.36	0.999970
0.58	0.122760	1.04	0.770434	1.49	0.976412	2.38	0.999976
0.60	0.135718	1.05	0.779794	1.50	0.977782	2.40	0.999980
0.61	0.149229	1.06	0.788860	1.52	0.980310	2.42	0.999984
0.62	0.163225	1.07	0.797636	1.54	0.982578	2.44	0.999987
0.63	0.177753	1.08	0.806128	1.56	0.984610	2.46	0.999989
0.64	0.192677	1.09	0.814342	1.58	0.986426	2.48	0.999991
0.65	0.207987	1.10	0.822282	1.60	0.988048	2.58	0.999 9925
0.66	0.223637	1.11	0.829950	1.62	0.989492	2.55	0.999 9956
0.67	0.239582	1.12	0.837356	1.64	0.990777	2.60	0.999 9974
0.68	0.255780	1.13	0.844502	1.66	0.991917	2.65	0.999 9984
0.69	0.272189	1.14	0.851394	1.68	0.992928	2.70	0.999 9990
0.70	0.288765	1.15	0.858038	1.70	0.993823	2.80	0.999 9997
0.71	0.305471	1.16	0.864442	1.72	0.994612	2.90	0.999 99990
0.72	0.322265	1.17	0.870612	1.74	0.995309	3.00	0.999 99997

> KS probabilities are invalid when the model parameters are estimated from the data. Some astronomers use them incorrectly.
>
> – Lillifors (1964)

Multivariate Case

Example – Paul B. Simpson (1951)

$F(x, y) = ax^2y + (1-a)y^2x, \qquad 0 < x, y < 1$

$(X_1, Y_1) \sim F.$ $\qquad F_1$ denotes the EDF of (X_1, Y_1)

$$\begin{aligned} P(|F_1(x,y) - F(x,y)| < .72, \text{ for all } x, y) \\ > .065 \text{ if } a = 0, \quad (F(x,y) = y^2x) \\ < .058 \text{ if } a = .5, \quad (F(x,y) = \frac{1}{2}xy(x+y)) \end{aligned}$$

Numerical Recipe's treatment of a 2-dim KS test is mathematically invalid.

Processes with estimated parameters

$\{F(.;\theta) : \theta \in \Theta\}$ – a family of continuous distributions

Θ is a open region in a p-dimensional space.

$X_1, \dots, X_n$ sample from F

Test $F = F(.;\theta)$ for some $\theta = \theta_0$

Kolmogorov-Smirnov, Cramér-von Mises statistics, etc., when θ is estimated from the data, are continuous functionals of the empirical process

$$Y_n(x;\hat{\theta}_n) = \sqrt{n}\big(F_n(x) - F(x;\hat{\theta}_n)\big)$$

$\hat{\theta}_n = \theta_n(X_1, \dots, X_n)$ is an estimator θ

F_n – the EDF of $X_1, \dots, X_n$

Bootstrap

G_n is an estimator of F, based $X_1, \ldots, X_n$.

$X_1^*, \ldots, X_n^*$ i.i.d. from G_n

$\hat{\theta}_n^* = \theta_n(X_1^*, \ldots, X_n^*)$

$F(.;\theta)$ is Gaussian with $\theta = (\mu, \sigma^2)$

If $\hat{\theta}_n = (\bar{X}_n, s_n^2)$, then

$\hat{\theta}_n^* = (\bar{X}_n^*, s_n^{*2})$

Parametric bootstrap if $G_n = F(.;\hat{\theta}_n)$

$X_1^*, \ldots, X_n^*$ i.i.d. $F(.;\hat{\theta}_n)$

Nonparametric bootstrap if $G_n = F_n$ (EDF)

Parametric bootstrap

$X_1^*, \ldots, X_n^*$ sample generated from $F(.; \hat{\theta}_n)$

In Gaussian case $\hat{\theta}_n^* = (\bar{X}_n^*, s_n^{*2})$.

Both

$$\sqrt{n} \sup_x |F_n(x) - F(x; \hat{\theta}_n)|$$

and

$$\sqrt{n} \sup_x |F_n^*(x) - F(x; \hat{\theta}_n^*)|$$

have the same limiting distribution

In XSPEC package, the parametric bootstrap is command FAKEIT, which makes Monte Carlo simulation of specified spectral model

Nonparametric bootstrap

$X_1^*, \ldots, X_n^*$ sample from F_n
i.e., a simple random sample from $X_1, \ldots, X_n$.

Bias correction

$$B_n(x) = \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n))$$

is needed.

Both

$$\sqrt{n} \sup_x |F_n(x) - F(x; \hat{\theta}_n)|$$

and

$$\sup_x |\sqrt{n} \left(F_n^*(x) - F(x; \hat{\theta}_n^*)\right) - B_n(x)|$$

have the same limiting distribution.

XSPEC does not provide a nonparametric bootstrap capability

χ^2 **type statistics** – (Babu, 1984, Statistics with linear combinations of chi-squares as weak limit. *Sankhyā*, Series A, **46**, 85-93.)

U-**statistics** – (Arcones and Giné, 1992, On the bootstrap of U and V statistics. *The Ann. of Statist.*, **20**, 655–674.)

Confidence limits under misspecification

$X_1, \ldots, X_n$ data from unknown H.

H may or may not belong to the family $\{F(.;\theta) : \theta \in \Theta\}$

H is closest to $F(., \theta_0)$

Kullback-Leibler information

$\int h(x) \log \big(h(x)/f(x;\theta)\big) d\nu(x) \geq 0$

$\int |\log h(x)| h(x) d\nu(x) < \infty$

$\int h(x) \log f(x;\theta_0) d\nu(x) = \max_{\theta\in\Theta} \int h(x) \log f(x;\theta) d\nu(x)$

For any $0 < \alpha < 1$,

$$P\Big(\sqrt{n}\sup_x |F_n(x) - F(x;\hat{\theta}_n) - (H(x) - F(x;\theta_0))| \leq C^*_\alpha\Big) \; - \alpha \to 0$$

C^*_α is the α-th quantile of

$$\sup_x |\sqrt{n}\,\big(F^*_n(x) - F(x;\hat{\theta}^*_n)\big) - \sqrt{n}\big(F_n(x) - F(x;\hat{\theta}_n)\big)|$$

This provide an estimate of the distance between the true distribution and the family of distributions under consideration.

Similar conclusions can be drawn for von Mises-type distances

$$\int \big(F_n(x) - F(x;\hat{\theta}_n) - (H(x) - F(x;\theta_0))\big)^2 \, dF(x;\theta_0),$$

$$\int \big(F_n(x) - F(x;\hat{\theta}_n) - (H(x) - F(x;\theta_0))\big)^2 \, dF(x;\hat{\theta}_n).$$

References

- Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory*, (B. N. Petrov and F. Csaki, Eds). Akademia Kiado, Budapest, 267-281.

- Babu, G. J., and Bose, A. (1988). Bootstrap confidence intervals. *Statistics & Probability Letters*, **7**, 151-160.

- Babu, G. J., and Rao, C. R. (1993). Bootstrap methodology. In *Computational statistics*, Handbook of Statistics **9**, C. R. Rao (Ed.), North-Holland, Amsterdam, 627-659.

- Babu, G. J., and Rao, C. R. (2003). Confidence limits to the distance of the true distribution from a misspecified family by bootstrap. *J. Statistical Planning and Inference*, **115**, no. 2, 471-478.

- Babu, G. J., and Rao, C. R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhyā*, **66**, no. 1, 63-74.

- Getman, K. V., and 23 others (2005). Chandra Orion Ultradeep Project: Observations and source lists. *Astrophys. J. Suppl.*, **160**, 319-352.