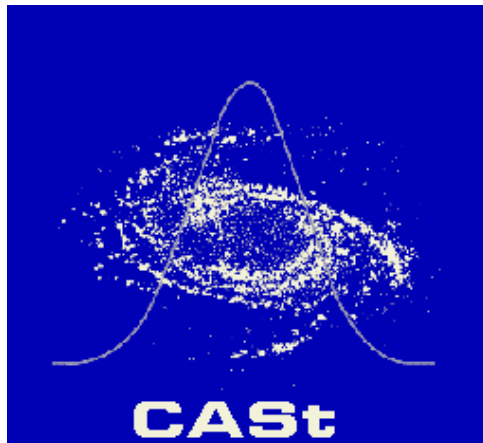


Goodness of Fit

G. Jogesh Babu

Center for Astrostatistics

<http://astrostatistics.psu.edu>



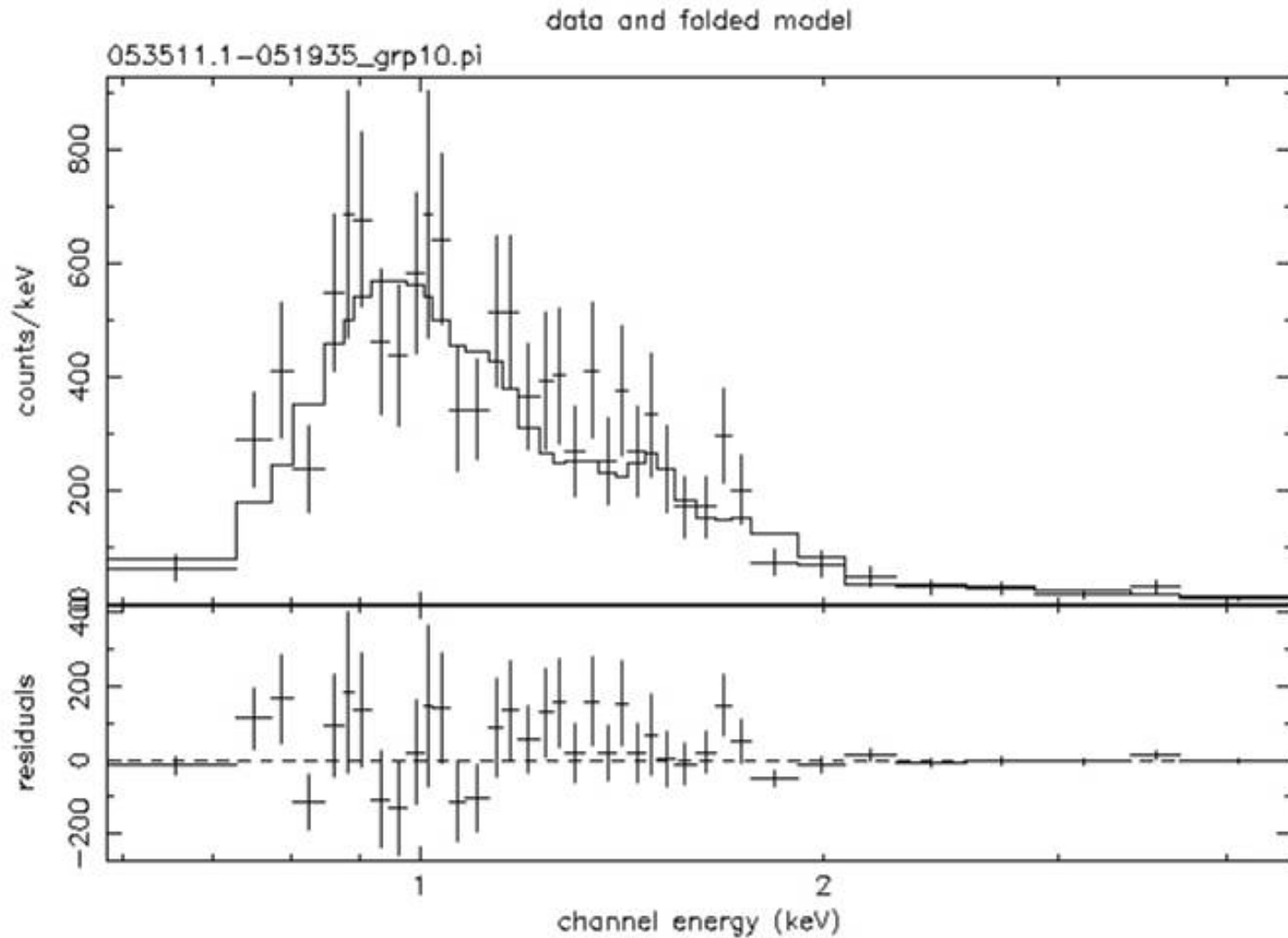
VOSTAT

grist

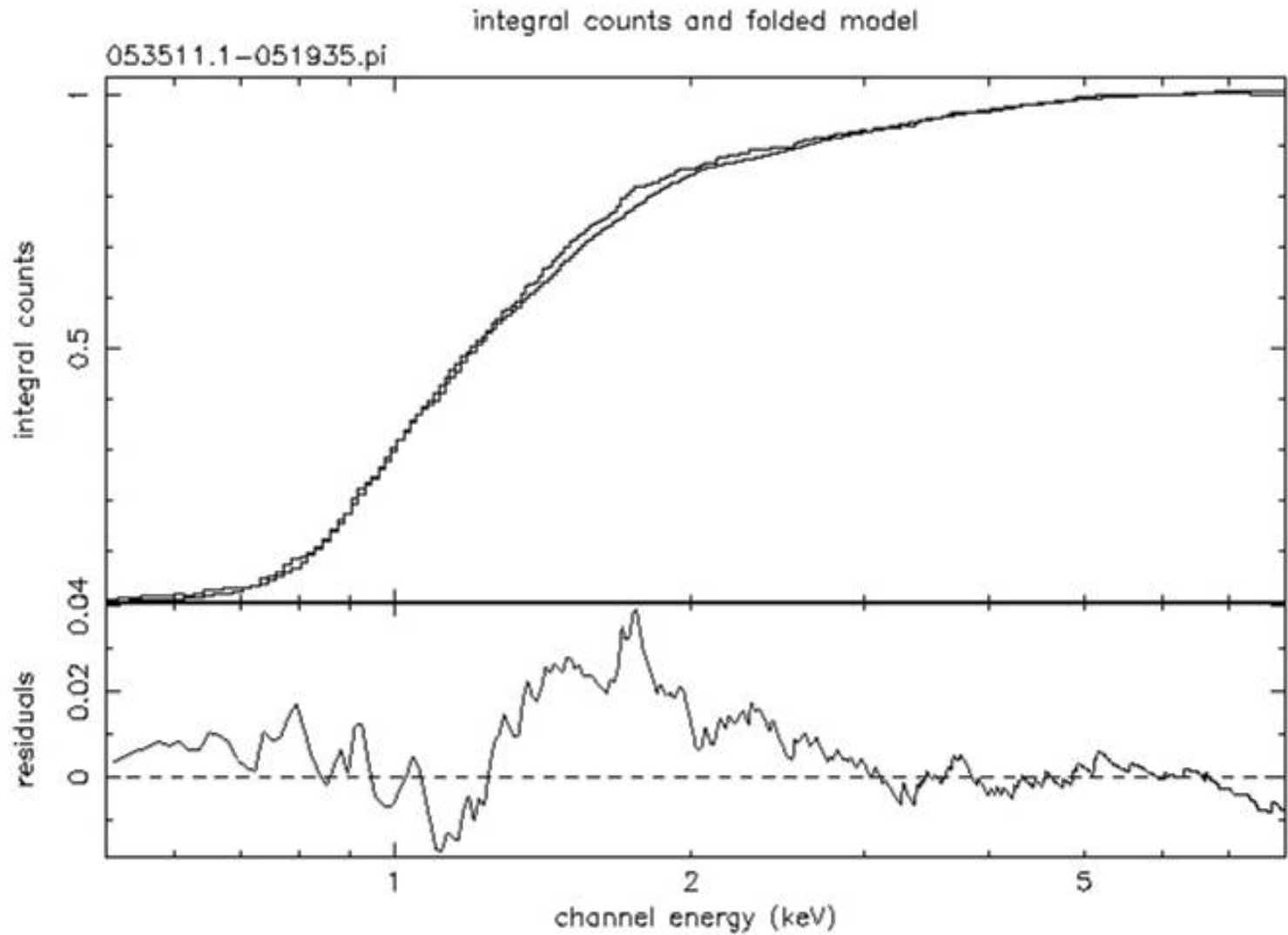
Astrophysical Inference from Data

Fitting Astrophysical data

- Non-linear regression
- Density (shape) estimation
- Parametric modeling
- Goodness of fit

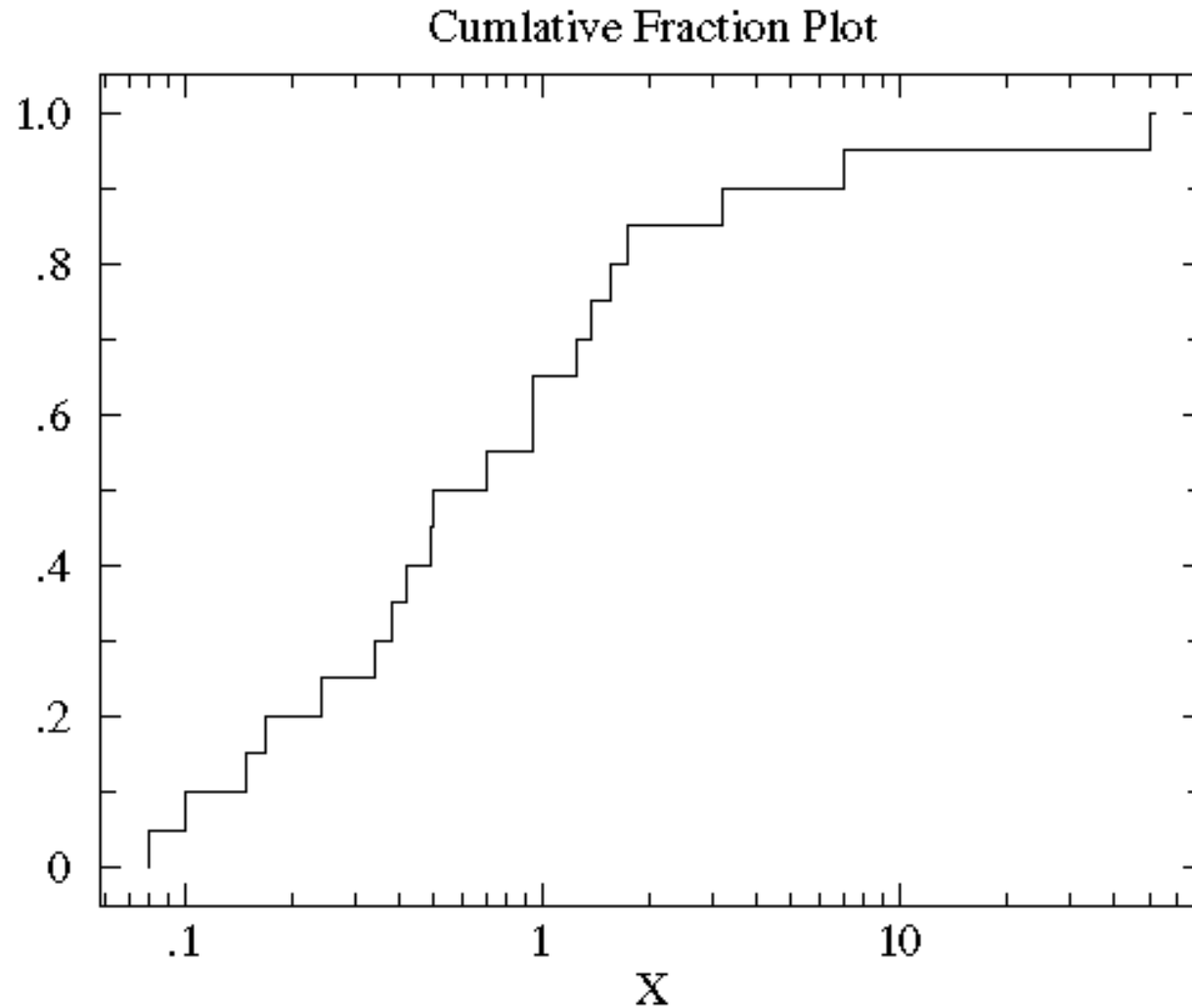


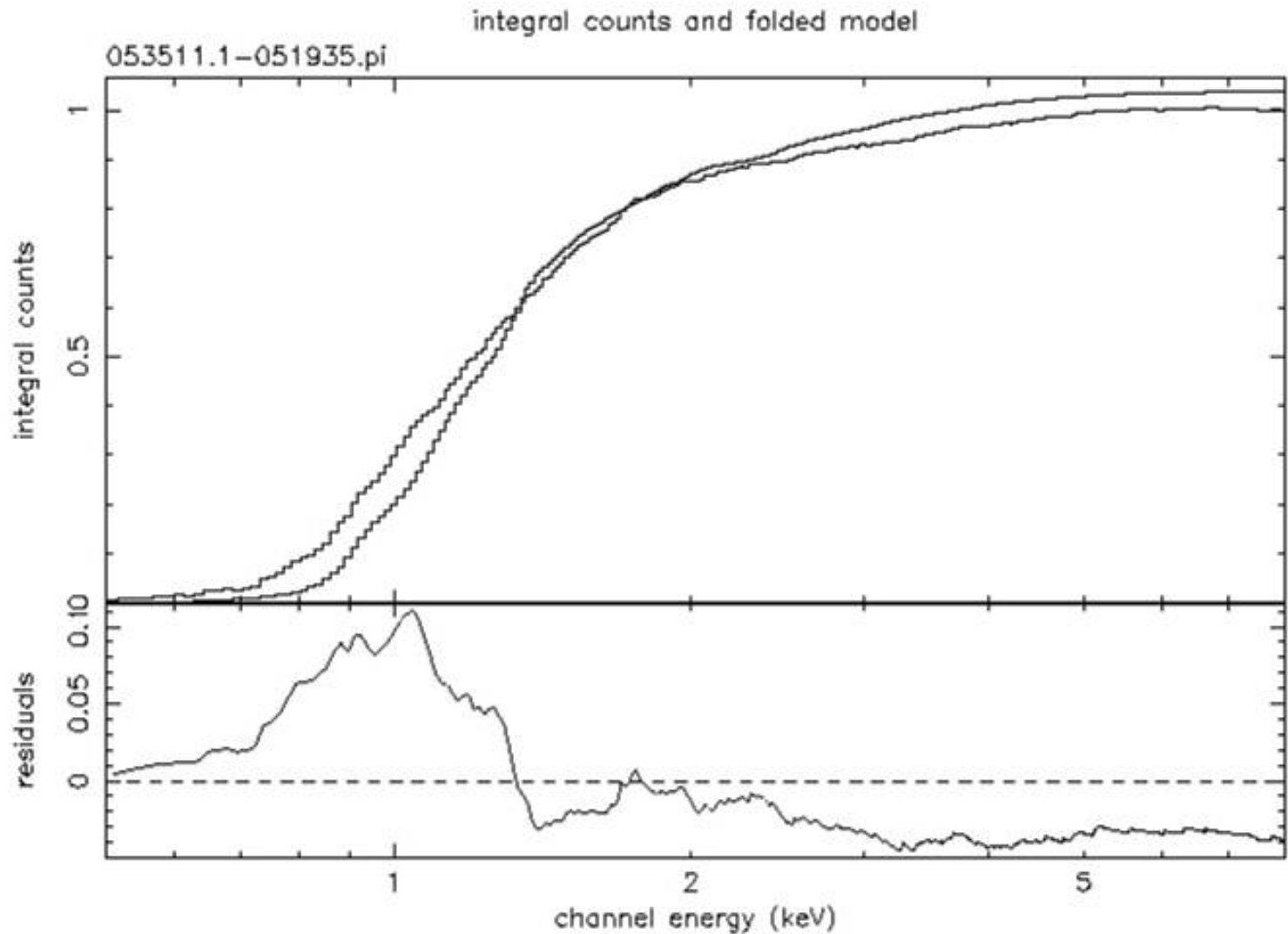
Chandra X-ray Observatory ACIS data
COUP source # 410 in Orion Nebula with 468 photons
Fitting to binned data using χ^2 (XSPEC package)
Thermal model with absorption, $A_V \sim 1$ mag



Fitting to unbinned EDF
Maximum likelihood (C-statistic)
Thermal model with absorption

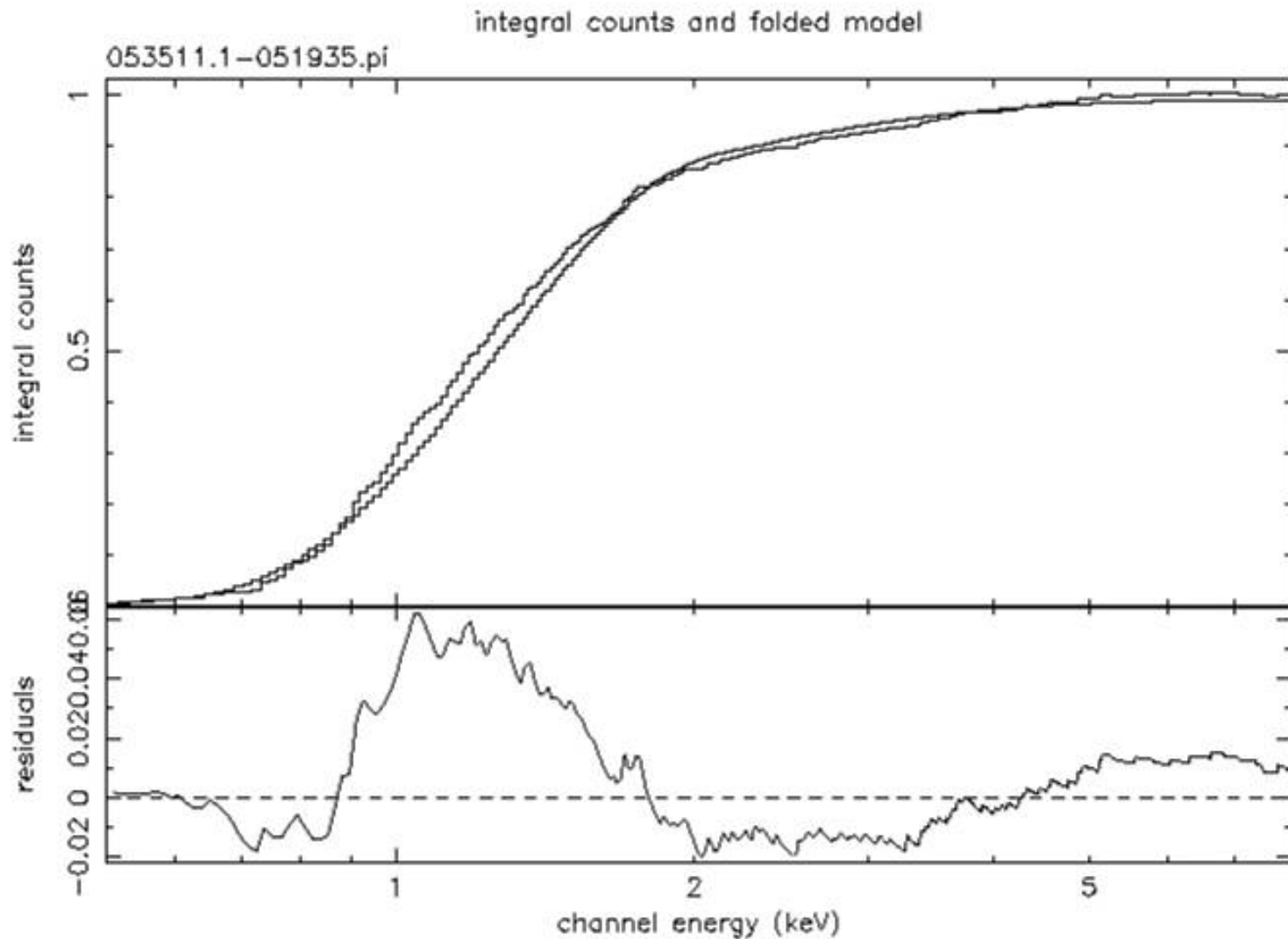
Empirical Distribution Function





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

Question : What is the 99% confidence interval for A_V ?

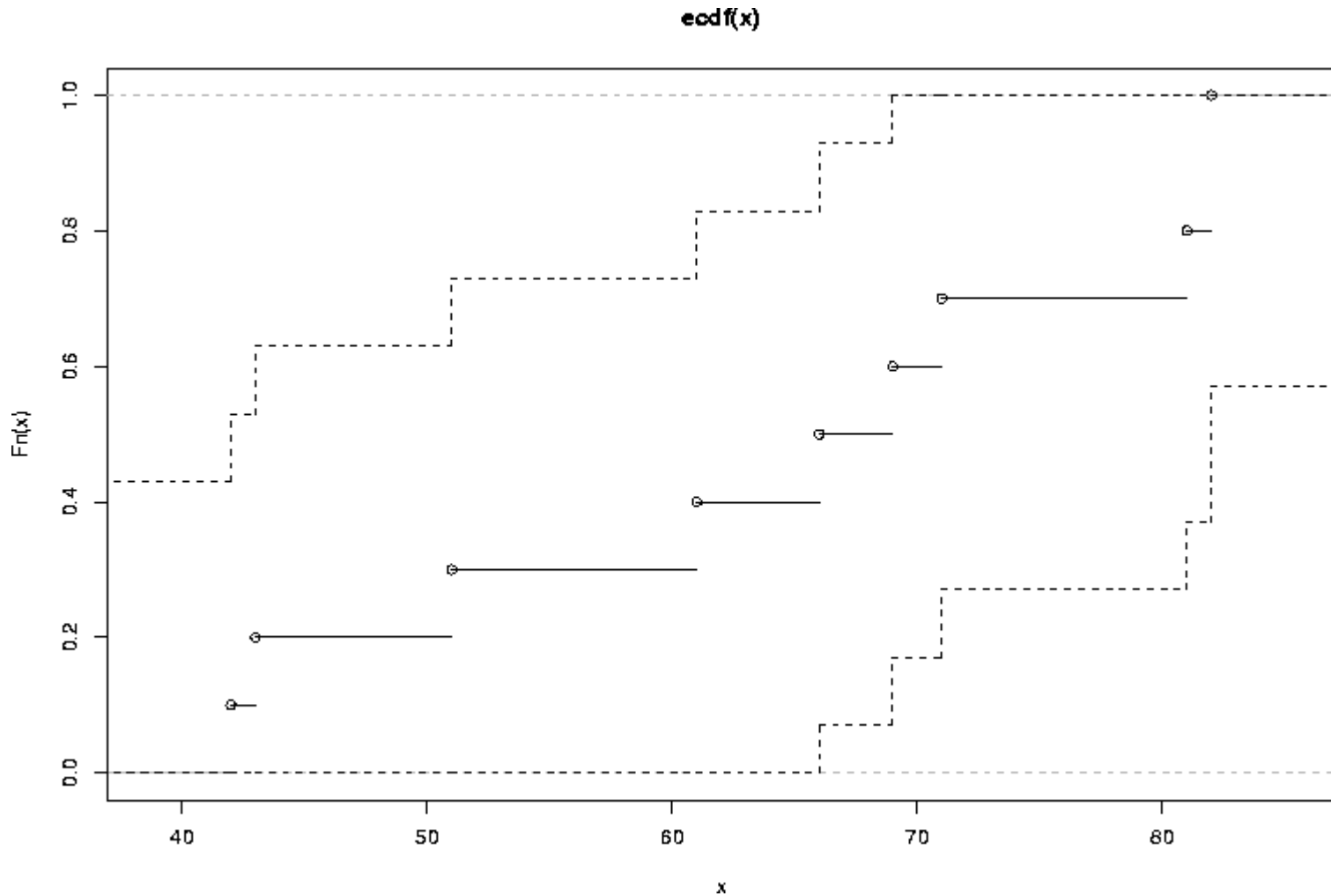


Incorrect model family

Power law model, absorption $A_V \sim 1$ mag

Question : Can a power law model be excluded with 99% confidence?

K-S Confidence bands



Model fitting

Aim: To provide most parsimonious `best' fit

To answer:

- Is the underlying nature of a 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?

- These issues arise when data are used to repudiate or support astrophysical theories but the underlying processes generating the data are not confidently known.
- We have developed nonparametric resampling methods for inference, when the data come from an unknown distribution which may or may not belong to a specified family

Statistics Based on EDF

Kolmogorov-Smirnov: $\text{Sup}_x |F_n(x) - F(x)|,$
 $\text{Sup}_x (F_n(x) - F(x))^+, \quad \text{Sup}_x (F_n(x) - F(x))^-$

Cramer - van Mises: $\int (F_n(x) - F(x))^2 dF(x)$

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

All of these statistics are distribution free

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	0.778794	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	0.9999925
0.66	0.223637	1.11	0.829950	1.62	0.989492	2.55	0.9999956
0.67	0.239582	1.12	0.837356	1.64	0.990777	2.60	0.9999974
0.68	0.255780	1.13	0.844502	1.66	0.991917	2.65	0.9999984
0.69	0.272189	1.14	0.851394	1.68	0.992928	2.70	0.9999990
0.70	0.288765	1.15	0.858038	1.70	0.993823	2.80	0.9999997
0.71	0.305471	1.16	0.864442	1.72	0.994612	2.90	0.99999990
0.72	0.322265	1.17	0.870612	1.74	0.995309	3.00	0.99999997

Statistics Based on EDF

Kolmogorov-Smirnov: $\text{Sup}_x |F_n(x) - F(x)|$

Cramer - van Mises: $\int (F_n(x) - F(x))^2 dF(x)$

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

All of these statistics are no longer distribution free if parameters are estimated or the data is multivariate.

Multivariate Case

Warning: K-S does not work in multidimensions

Example – Paul B. Simpson (1951)

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

(X_1, Y_1) data from F , F_1 EDF of (X_1, Y_1)

$P(|F_1(x,y) - F(x,y)| < 0.72, \text{ for all } x, y)$ is

$$> 0.065 \text{ if } a = 0, \quad (F(x,y) = y^2 x)$$

$$< 0.058 \text{ if } a = 0.5, \quad (F(x,y) = xy(x+y)/2)$$

Processes with estimated Parameters

$\{F(\cdot; \theta): \theta \in \Theta\}$ - a family of distributions

X_1, \dots, X_n sample from F

Kolmogorov-Smirnov, Cramer-von Mises etc.,

when θ is estimated from the data, are

Continuous functionals of the empirical process

$$Y_n(\mathbf{x}; \theta_n) = \sqrt{n} (F_n(\mathbf{x}) - F(\mathbf{x}; \theta_n))$$

In the Gaussian case,

$$\theta = (\mu, \sigma^2) \text{ and } \theta_n = (\bar{X}, s_n^2)$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- **Darling, D. A.** (1955). The Cramér-Smirnov test in the parametric case. *Ann. Math. Statist.*, **26**, 1–20.
- **Kac, M., Kiefer, J., and Wolfowitz, J.** (1955). On tests of normality and other tests of goodness of fit based on distance methods. *Ann. Math. Statist.*, **26**, 189–211.
- **Durbin, J.** (1973). Weak convergence of the sample distribution function when parameters are estimated. *Ann. of Statist.*, **1**, 279–290.

Bootstrap

\hat{F}_n is an estimator of F , based X_1, \dots, X_n .

X_1^*, \dots, X_n^* i.i.d. from \hat{F}_n

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

$F(\cdot; \theta)$ 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 $\hat{F}_n = F(\cdot; \hat{\theta}_n)$

X_1^*, \dots, X_n^* i.i.d. $F(\cdot; \mu, \sigma^2)$

Nonparametric bootstrap if $\hat{F}_n = F_n$

Parametric Bootstrap

X_1^*, \dots, X_n^* sample generated from $F(\cdot; \theta_n)$.

In Gaussian case $\theta_n^* = (\bar{X}_n^*, S_n^{*2})$.

Both $\sqrt{n} \sup_x |F_n(x) - F(x; \theta_n)|$ and

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

have the same limiting distribution

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

$$\begin{aligned} Y_n^P(x) &= \sqrt{n}(F_n^*(x) - F(x; \hat{\theta}_n^*)) \\ &= \sqrt{n}(F_n^*(x) - F(x; \hat{\theta}_n) + F(x; \hat{\theta}_n) - F(x; \hat{\theta}_n^*)) \end{aligned}$$

and the sample process

$$\begin{aligned} Y_n(x) &= \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n)) \\ &= \sqrt{n}(F_n(x) - F(x; \theta) + F(x; \hat{\theta}) - F(x; \hat{\theta}_n)) \end{aligned}$$

converge to the same Gaussian process Y

Nonparametric Bootstrap

X_1^*, \dots, X_n^* *i.i.d.* from F_n .

A bias correction

$$B_n(x) = F_n(x) - F(x; \theta_n)$$

is needed.

$$\sqrt{n} \sup_x |F_n(x) - F(x; \theta_n)| \text{ and}$$

$$\sqrt{n} \sup_x |F_n^*(x) - F(x; \theta_n^*) - B_n(x)|$$

have the same limiting distribution

(XSPEC does not provide a nonparametric bootstrap capability)

$$\begin{aligned}
Y_n^N(x) &= \sqrt{n}(F_n^*(x) - F(x; \hat{\theta}_n^*)) - B_n(x) \\
&= \sqrt{n}(F_n^*(x) - F_n(x) + F(x; \hat{\theta}_n) - F(x; \hat{\theta}_n^*))
\end{aligned}$$

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

Y_n, Y_n^N converge to the same Gaussian process Y

$$\begin{aligned}
Y_n(x) &= \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n)) \\
&= \sqrt{n}(F_n(x) - F(x; \theta) + F(x; \hat{\theta}) - F(x; \hat{\theta}_n))
\end{aligned}$$

- **Chi-Square type statistics** – (Babu, 1984, Statistics with linear combinations of chi-squares as weak limit. *Sankhya*, Series A, **46**, 85-93.)
- **U-statistics** – (Arcones and Giné, 1992, On the bootstrap of U and V statistics. *Ann. of Statist.*, **20**, 655–674.) Resampling methods for semi-parametric goodness-of-fit tests – p. 12/28

Confidence limits under misspecification of model family

X_1, \dots, X_n data from unknown H .

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

H is closest to $F(\cdot; \theta_0)$, in Kullback - Leibler information

$$\int h(x) \log (h(x)/f(x; \theta)) dv(x) \geq 0$$

$$\int h(x) |\log (h(x))| dv(x) < \infty$$

$$\int h(x) \log f(x; \theta_0) dv(x) = \max_{\theta} \int h(x) \log f(x; \theta) dv(x)$$

- **White, H.** (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.
- **Nishii, R.** (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multi. Analysis.*, **27**, 392–403.
- **Foutz, R. V., and Srivastava, R. C.** (1977). The performance of the likelihood ratio test when the model is incorrect. *Ann. Statist.*, **5**, 1183–1194.

$$U_n(x; \hat{\theta}_n) = \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n)) - \sqrt{n}(H(x) - F(x; \theta_0))$$

converges weakly to a Gaussian process U

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

and U_n converge to the same Gaussian process.

The results are established under the same conditions as before.

If $m \rightarrow \infty$ and $m/n \rightarrow 0$, then U_n and

$$Y_n^m(x) = \sqrt{m}(F_m^*(x) - F(x; \hat{\theta}_m^*))$$

converge to the same limiting Gaussian Process. This can be used to obtain confidence bands for $H - F(\cdot; \theta_0)$.

For any $0 < \alpha < 1$,

$$P(\sqrt{n} \sup_x |F_n(x) - F(x; \theta_n) - (H(x) - F(x; \theta_0))| < C_{\alpha}^*) \rightarrow \alpha$$

C_{α}^* is the α -th quantile of

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

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 statistics

$$\int (F_n(x) - F(x; \theta_n) - (H(x) - F(x; \theta_0)))^2 dF(x; \theta_0)$$

$\Lambda \subset \Theta$ is a closed nbhd of θ_0

(A) $g(x; \theta) = \frac{\partial}{\partial \theta} F(x; \theta)$ uniformly continuous in x and $\theta \in \Lambda$

(B) For some $\epsilon_n = \epsilon_n(X_1, \dots, X_n) \rightarrow_p 0$,

$$\hat{\theta}_n - \theta_0 = \frac{1}{n} \sum_{i=1}^n \ell(X_i; \theta_0) + \frac{1}{\sqrt{n}} \epsilon_n$$

(P) For some $\epsilon_n^* \rightarrow 0$ in probability under the bootstrap measure,

$$\hat{\theta}_n^* - \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n \ell(X_i^*; \hat{\theta}_n) + \frac{1}{\sqrt{n}} \epsilon_n^*$$

(N) For some $\epsilon_n^* \rightarrow 0$ in probability under the bootstrap measure,

$$\hat{\theta}_n^* - \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n \ell(X_i^*; \theta_0) - \frac{1}{n} \sum_{i=1}^n \ell(X_i; \theta_0) + \frac{1}{\sqrt{n}} \epsilon_n^*$$

(L1) For $\theta \in \Lambda$, $L(\theta) = \int \ell'(x; \theta) \ell(x; \theta) dF(x; \theta)$ is a finite non-negative definite matrix.

(L2) As $\gamma \rightarrow \infty$, (Lindberg type condition)

$$\sup_{\theta \in \Lambda} \int_{\{\|\ell(x; \theta)\| > \gamma\}} \|\ell(x; \theta)\|^2 dF(x; \theta) \rightarrow 0$$

(L3) For all x , $h(x; \theta) = \int_{-\infty}^x \ell(t; \theta) dF(t; \theta)$ is continuous at θ_0 .

These conditions are needed for the convergence of $\sqrt{n}(\theta_n - \theta)$

The Covariance Function

The covariance function of Y is given by

$$\begin{aligned} \text{Cov}(Y(x), Y(y)) &= \min(F(x; \theta_0), F(y; \theta_0)) - F(x; \theta_0)F(y; \theta_0) \\ &\quad - h(x; \theta_0)g(y; \theta_0)' - h(y; \theta_0)g(x; \theta_0)' \\ &\quad + g(x; \theta_0)L(\theta_0)g(y; \theta_0)' \end{aligned}$$

For the sample median $\hat{\theta}_n$,

$$\ell(x; \theta) = \frac{1}{f(\theta)} \left(I_{\{x \leq \theta\}} - \frac{1}{2} \right), \quad f \text{ density of } F$$

For L -statistic, $\hat{\theta}_n = \int x \omega(F_n(x)) dF_n(x)$,

$$\ell(y; \theta) = \int_{-\infty}^{\infty} (F(x; \theta) - I_{\{y \leq x\}}) \omega(F(x; \theta)) dx,$$

ω satisfies Lipschitz condition of order 1 on each of the intervals

$$(a_{i-1}, a_i), \quad i = 1, \dots, k+1, \quad a_0 = 0 < a_1 < \dots < a_k < a_{k+1} = 1$$

F^{-1} is continuous at a_1, \dots, a_k

Babu & Singh (1984). Asymptotic representations related to jackknifing and bootstrapping L-statistics. *Sankhyā*, A 46, 195-206.

$$(B) \quad \hat{\theta}_n - \theta_0 = \frac{1}{n} \sum_{i=1}^n \ell(X_i; \theta_0) + \frac{1}{\sqrt{n}} \epsilon_n$$

$$(N) \quad \hat{\theta}_n^* - \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n \ell(X_i^*; \theta_0) - \frac{1}{n} \sum_{i=1}^n \ell(X_i; \theta_0) + \frac{1}{\sqrt{n}} \epsilon_n^*$$

The bias corrected process

$$\begin{aligned} Y_n^N(x) &= \sqrt{n} (F_n^*(x) - F(x; \hat{\theta}_n^*)) - B_n(x) \\ &= \sqrt{n} (F_n^*(x) - F_n(x) + F(x; \hat{\theta}_n) - F(x; \hat{\theta}_n^*)) \\ W_n^*(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(I_{\{X_i^* \leq x\}} - F_n(x) \right. \\ &\quad \left. - \left(\ell(X_i^*; \theta_0) - \frac{1}{n} \sum_{i=1}^n \ell(X_i; \theta_0) \right) g(x; \theta_0)' \right) \end{aligned}$$

$\sup_x |Y_n^N(x) - W_n^*(x)| \rightarrow 0$ in the bootstrap measure
for almost all sample sequences X_1, \dots, X_n

References

- **Bootstrap Methodology**

G. J. Babu and C. R. Rao (1993), Handbook of Statistics, Vol 9, Chapter 19.

- **Bootstrap Techniques for Signal Processing**

Abdelhak M. Zoubir and D. Robert Iskander, Cambridge, U.K.: Cambridge University Press, 2004.

This book serves as a handbook on *bootstrap* for engineers, to analyze complicated data with little or no model assumptions.

Bootstrap has found many applications in engineering field including, artificial neural networks, biomedical engineering, environmental engineering, image processing, and Radar and sonar signal processing.

Majority of the applications are taken from signal processing literature.

Summary

- MLE and Chi-square fits are often OK, but it is hard to evaluate goodness of fit
- KS probabilities often wrong (multivariate, estimated parameters)
- Use instead parametric bootstrap to get parametric confidence bands and evaluate different model families