SUMMER SCHOOL IN STATISTICS FOR ASTRONOMERS & PHYSICISTS-2

Statistical Inference for Astronomers

MODEL SELECTION AND EVALUATION

Goodness of fit and likelihood ratio tests

C.R,Rao
Statistics Department
Pennsylvania State University

June 8 , 2006

FISHERIAN FRAMEWORK

(1922) On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc.* **A222**, 309-368.

THREE METHODOLOGICAL PROBLEMS OF STATISTICS

- **Specification**: selection of a model (family of probability distributions) applicable to observed data

  How to choose a model?

  Pearson-Fisher controversy, (χ^2, ρ) of Pearson

  H.F. Inman. *Am. Statistician*, **48**, 2-11 (1994)

- **Estimation**: concepts of consistency, efficiency, asymptotic variance, sufficiency and information. (Maximum likelihood)

- **Testing of null hypothesis**: choice of a test statistic? Interpretation of p-values. Purpose of a test.

GOODNESS OF FIT TESTS

Karl Pearson Chi-Square (1900)

The Holy Trinity

Neyman Pearson likelihood ratio test (1928)
Wald test (1948)
Rao's Score test (1948)

Other tests

Kolmogorov- Smirnoff test
Cramer- von Mises test

A general theorem

References:

E. Lehmann(1998). *Elements of Large Sample Theory* Springer, pages 526-535.

C.R.Rao (1973). *Linear Statistical Inference and its Applications*, John Wiley, Chapter 6.

INFORMATION THEORETIC CRITERIA FOR MODEL SELECTION

AIC, AICc, GIC, QAIC, QAICc
TIC, WIC, BIC

OTHERS

NIC, Mallow's Cp

CROSS VALIDATION FOR MODEL ACCURACY

Training Sample
Test Sample
Bootstrap Sample

References:

C.R.Rao and Y.Wu (2000). On model selection, In *Model Selection,* Ed. Lahiri, IMS Lecture Notes, 38, 1-64.

K.P.Burnham and D.R. Anderson (1998). *Model Selection and Inference, A Practical Information Theoretic Approach*, Springer

BOOTSTRAP METHODOLOGY

F is a distribution function DF.

Sample: $\tilde{x} = (x_1, \ldots, x_n)$, i.i.d. random variables.

Sample statistics: $t(\tilde{x})$, estimator or test criterion.

Distribution of $t(\tilde{x})$.

Draw independent samples of size n, if F is known

$$\tilde{x}_1, \tilde{x}_2, \ldots$$

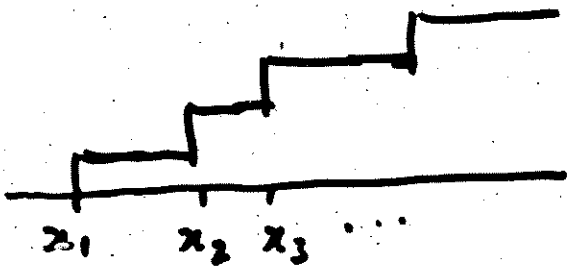
and compute

$$t(\tilde{x}_1), t(\tilde{x}_2), \ldots, t(\tilde{x}_N), \ldots$$

The sequence characterizes the sampling distribution F_t of $t(\tilde{x})$. By choosing N sufficiently large, we can estimate the quantiles of the distribution of $T(\tilde{x})$ with any desired accuracy.

If F is not known, we may proceed as follows
Estimate F using the sample $\tilde{x} = (x_1, \ldots, x_n)$

$$\hat{F}: \begin{cases} \text{Value} & x_1, \ldots, x_n \\ \text{Probability} & \frac{1}{n}, \ldots, \frac{1}{n} \end{cases}$$



Bootstrap distribution of $t(\tilde{x})$

Draw samples $\tilde{x}_1^*, \tilde{x}_2^*, \ldots, \tilde{x}_N^*, \ldots$ from $\hat{F}(x)$ and compute

$$t(\tilde{x}_1^*), t(\tilde{x}_2^*), \ldots, t(\tilde{x}_N^*), \ldots$$

The distribution F_t^* defined by the sequence is called *the bootstrap distribution of* $t(\tilde{x})$.

Reference: Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. Springer

PRELIMINARIES

- Probability model for a random variable X

Probability density at $x = x$: $f(x, \theta)$, X continuous
Probability of event $X = x$: $f(x, \theta)$, X discrete

- Sample of size n

$$S : \underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$$

n independent observations drawn from $f(x, \theta)$.

- Likelihood and log likelihood of $\theta = (\theta_1, \ldots, \theta_q)$

$$L(\theta|S) = f(x_1, \theta) \ldots f(x_n, \theta)$$
$$l(\theta|S) = \log L(\theta|x) = \log f(x_1, \theta) + \ldots + \log f(x_n, \theta)$$

- Score function is a q-vector

$$s = (s_1, \ldots, s_q)$$
$$s_i = \frac{\partial l}{\partial \theta_i} = \sum_{j=1}^{n} \frac{1}{f(x_j, \theta)} \frac{\partial f(x_j, \theta)}{\partial \theta_i}, i = 1, \ldots, q$$

- Maximum likelihood estimate

$$\hat{\theta} = \arg\max_{\theta} l(\theta|S)$$

usually obtained as some root of the equations (why?)

$$s_i = 0, \ i = 1, \ldots, n.$$

Karl Pearson Chi-Square Test

The dawn of statistical inference

In an article entitled, *Trial by Number*, Hacking (1984) says that the goodness-of-fit chi-square test introduced by Karl Pearson (1900), "ushered in a new kind of decision making" and gives it a place among the top 20 discoveries since 1900 considering all branches of science and technology. R.A. Fisher, who was involved in bitter controversies with Pearson, was appreciative of the chi-square test. In his book on *Statistical Methods for Research Workers* (1958, 13th edition, p.22), Fisher says, "This (chi-square), I believe is the great contribution to statistical methodology which the unsurpassed energy of Professor Pearson's work will be remembered," and devoted one full chapter on numerous ingenious applications of the chi-square test.

Pearson's chi-square is ideally applicable to qualitative data with a finite number, say s, of natural categories and the data are in the form of frequencies of individuals in different categories. The specified hypothesis is of the form

$$\pi_i = \pi_i(\theta),\ i = 1, \ldots, s$$

where the probability π_i in category i is a given function of a k-vector parameter θ.

Class Interval (bins)	observed frequency (O)	Expected frequency (E)
$a_1 - a_2$	O_1	$n\pi_1(\hat{\theta})$
$a_2 - a_3$	O_2	$n\pi_2(\hat{\theta})$
$\vdots$	$\vdots$	$\vdots$
$a_n - a_{s+1}$	O_s	$n\pi_s(\hat{\theta})$
Total	n	n

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \quad d.f = s - 1 - k$$

FITTING A BINOMIAL DISTRIBUTION

$p(r) = \binom{k}{r} \pi^r (1-\pi)^{k-r}$, k trials, r successes

$r = 0, \ldots, k$, $k+1$ classes

FREQUENCY DISTRIBUTION OF NUMBER OF BOYS IN FAMILIES OF SIZE EIGHT

FIT OF BINOMIAL DISTRIBURION

Number of Boys.	Number of Families Observed.	Expected.	Excess (x).	$\frac{x^2}{m}$.
0	215	165·22	+ 49·78	14·998
1	1485	1401·69	+ 83·31	4·952
2	5331	5202·65	+128·35	3·166
3	10649	11034·65	−385·65	13·478
4	14959	14627·60	+331·40	7·508
5	11929	12409·87	−480·87	18·633
6	6678	6580·24	+ 97·76	1·452
7	2092	1993·78	+ 98·22	4·839
8	342	264·30	+ 77·70	22·843
	53680	53680·00		91·869

Estimate of $\pi = .61$

$$\chi^2 = \sum_0^8 \frac{(O_i - E_i)^2}{E_i} = 91.869$$

d.f. = no. of classes (bins) − 1 − no. of parameters estimated

$= 9 - 1 - 1 = 7$

$P < .005$

FITTING A POISSON DISTRIBUTION

Prob for r events

$$e^{-\mu}\frac{\mu^r}{r!}, \quad r = 0, 1, \ldots$$

FREQUENCY DISTRIBUTION OF DEATHS DUE TO HORSE KICKS

(RECORD OF 10 ARMY CORPS OVER 20 YEARS)

FIT OF POISSON DISTRIBUTION

Deaths.	Frequency observed.	Expected.	χ^2
0	109	108·67	·0001
1	65	66·29	·0251
2	22	20·22	·1566
3	3	4·11	
4	1	·63	4.83 ·1426
5	…	·08	
6	…	·01	
	200	200	0·3244

(3 and 1 grouped: 4; 4·11, ·63, ·08, ·01 grouped: 4.83)

d·f = 2

$P > .85$

Estimate of $\mu = 0.61$

LIKELIHOOD RATIO TEST (LRT)

- Probability model

$$f(x,\theta),\ x \in R^p,\ \theta \in \Theta \subset R^q$$

- Sample

$$S = (x_1, \ldots, x_n),\ n\ i.i.d.'s$$

- Log likelihood

$$l(\theta|S) = \sum_1^n \log\ f(x_i,\theta).$$

TEST OF A SIMPLE HYPOTHESIS

$$H_{0s} : \theta = \theta_0 \text{ (specified)}$$

against the alternative

$$H_1 : \theta \neq \theta_0 \ (\theta \text{ unspecified})$$

LRT for H_0

$$2\left[l(\hat{\theta}|S) - l(\theta_0|S)\right] \sim \chi^2(q)$$

where $\hat{\theta}$ is the ML of θ and q is the dimension of the parameter θ.

TEST OF A COMPOSITE HYPOTHESIS

H_{0c} : θ belongs to a subset $\Theta_0 \subset \Theta$, especially defined by a set of k independent restrictions $g_1(\theta) = 0, \ldots, g_r(\theta) = 0$

$$H_1 : \theta \in \Theta - \Theta_0.$$

LRT for H_{0c}

$$2\left[l(\hat{\theta}|S) - l(\hat{\theta}_r|S\right] \sim \chi^2(r)$$

where

$$\hat{\theta} = \arg\max_{\theta \in \Theta} [l(\theta|S)], \text{ the full MLE}$$

$$\hat{\theta}_r = \arg \max_{g_1(\theta)=\ldots=g_k(\theta)=0} [l(\theta|S)]$$

$r =$ the number of independent restrictions.

Note: the null hypothesis is a subset of Θ, and is referred to as a nested set.

WALD TEST

Test of a Simple hypothesis

$$H_{0s} : \theta_0 \in \Theta \subset R^q$$

against the alternative

$$H_1 : \theta \neq \theta_0, \theta \in \Theta.$$

Let $\hat{\theta}$ is the MLE of θ. The Wald test for H_0 is

$$W_s = n\left(\hat{\theta} - \theta_0\right)' I\left(\hat{\theta}\right)\left(\hat{\theta} - \theta_0\right) \sim \chi^2(q)$$

where $I(\theta)$ is the information matrix. W_s is a quadratic form. For example if $q = 2$,

$$\hat{\theta} = (5,3),\ \theta_0 = (1,2),\ \text{specified}$$

$$I(\hat{\theta}) = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$$

$$W_s = (5,3)\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} 5 \\ 3 \end{pmatrix} = (13,14)\begin{pmatrix} 5 \\ 3 \end{pmatrix} = 107(2df)$$

Test of a composite hypothesis

$$H_0 : g_1(\theta) = \cdots = g_r(\theta) = 0.$$

Let

$$g(\theta) = (g_1(\theta), \ldots, g_r(\theta))'$$

$$M(\theta) = \begin{pmatrix} \frac{\partial g_1}{\partial \theta_1}, & \cdots & \frac{\partial g_1}{\partial \theta_q} \\ \cdot & \cdots & \cdot \\ \frac{\partial g_r}{\partial \theta_1}, & \cdots & \frac{\partial g_r}{\partial \theta_q} \end{pmatrix}$$

$$W_c = ng(\hat{\theta})'\left[M(\hat{\theta})I(\hat{\theta})^{-1}M(\hat{\theta})'\right]^{-1} g(\hat{\theta}) \sim \chi^2(r)$$

$$(r \times q)(q \times q)(q \times r).$$

The test is not invariant to transformation of H_0.

RAO'S SCORE TEST

Test of a simple hypothesis

$$H_0 : \theta = \theta_0$$

against the alternative

$$H_1 : \theta \neq \theta_0.$$

Recall the Score function which is q-vector

$$s(\theta) = \left(\frac{\partial l}{\partial \theta_1}, \ldots, \frac{\partial l}{\partial \theta_q} \right)'.$$

The score test for H_0 is

$$R_s = [s(\theta_0)]' \, [I(\theta_0)]^{-1} \, [s(\theta_0)] \sim \chi^2(q).$$

Note: The test does not involve the computation of $\hat{\theta}$, the MLE of θ unlike in likelihood ratio and Wald tests.

The score test for the composite hypothesis

$$H_0 : g_1(\theta) = \ldots = g_r(\theta) = 0$$
$$[r(\tilde{\theta})]'[I(\tilde{\theta})]^{-1}[r(\tilde{\theta})] \sim \chi^2(r)$$

where $\tilde{\theta}$ is the MLE under the restriction of H_0 i.e.

$$\tilde{\theta} = \arg \max_{g_1(\theta)=\ldots=g_r(\theta)=0} l(\theta|S).$$

Note: All the three tests of Holy Trinity are asymptotically equivalent.

Karl Pearson's chisquare tests is a special case of Rao's score test.

Reference may be made to Rao (1973) and Lehmann (1998) for some applications and comments on these tests.

All these tests are not applicable when MLE's do not exist and are not well behaved. Further, they may not be applicable when θ under null hypothesis is on the border of the admissible set Θ. Consider for instance the null hypothesis

$$H_0 : f(x|\theta) = N(x|\mu_1, \sigma_1^2)$$

and the alternative

$$H_1 : f(x|\theta) = \alpha_1 N(x|\mu_1, \sigma_1^2) + \alpha_2 N(x|\mu_2, \sigma_2^2)$$
$$\alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_1 + \alpha_2 = 1.$$

All the above tests are not applicable in this case.

A GENERAL THEOREM

Let $y_1, \ldots, y_n$ be random variables such that

$$y_i \sim N\left(g(x_i, \theta), \sigma_i^2\right)$$
$$i = 2, \ldots, n$$

where x_i are fixed covariates, θ is unknown p-vector parameter and σ_i^2 are known. The maximum likelihood estimate of θ is

$$\hat{\theta} = \arg\min_{\theta} \sum_{1}^{n} \left(\frac{y_i - g(x_i\theta)}{\sigma_i} \right)^2 .$$

The χ^2 goodness-of-fit is

$$\sum_{1}^{n} \left(\frac{y_i - g(x_i, \hat{\theta})}{\sigma_i} \right)^2 , \text{ d.f. } n - p$$

where p is the number of unknown parameters.

Table 2. XRT 2-10 keV flux

T(mid)[a] (s)	T(exp) (s)	Flux[b]	T(mid)[a] (s)	T(exp) (s)	Flux[b]
133	5.	122.7 ± 5.7	578	10.	29.8 ± 2.0
143	5.	109.5 ± 5.4	598	10.	28.5 ± 2.0
153	5.	101.4 ± 5.2	618	10.	29.1 ± 2.0
163	5.	92.0 ± 4.9	638	10.	24.8 ± 1.8
173	5.	86.8 ± 4.8	658	10.	27.3 ± 1.9
183	5.	83.7 ± 4.7	678	10.	24.6 ± 1.8
193	5.	77.2 ± 4.5	708	20.	24.2 ± 1.3
203	5.	69.4 ± 4.3	748	20.	20.4 ± 1.2
213	5.	69.2 ± 4.3	788	20.	19.8 ± 1.2
223	5.	62.4 ± 4.1	828	20.	19.0 ± 1.1
233	5.	65.0 ± 4.1	868	20.	16.3 ± 1.1
243	5.	57.2 ± 3.9	908	20.	18.5 ± 1.1
253	5.	54.6 ± 3.8	948	20.	17.6 ± 1.1
263	5.	54.3 ± 3.8	988	20.	16.3 ± 1.1
278	10.	50.8 ± 2.6	1028	20.	15.5 ± 1.4
298	10.	49.8 ± 2.6	6009	150.	1.072 ± 0.134
318	10.	45.4 ± 2.5	6309	150.	1.331 ± 0.153
338	10.	45.5 ± 2.5	6609	150.	0.987 ± 0.126
358	10.	46.9 ± 2.5	6909	150.	1.010 ± 0.152
378	10.	40.2 ± 2.3	11809	350.	0.560 ± 0.056
398	10.	41.7 ± 2.4	12509	350.	0.482 ± 0.052
418	10.	39.4 ± 2.3	18109	350.	0.331 ± 0.052
438	10.	39.9 ± 2.3	23859	2000.	0.173 ± 0.039
458	10.	34.8 ± 2.2	27859	2000.	0.108 ± 0.028
478	10.	31.9 ± 2.1	35859	2000.	0.079 ± 0.021
498	10.	31.5 ± 2.1	81459	5400.	0.0361 ± 0.0088
518	10.	32.5 ± 2.1	99102	11850.	0.0273 ± 0.0078
538	10.	27.4 ± 1.9	165485	22000.	0.0080 ± 0.0016
558	10.	32.4 ± 2.1	412515	42000.	0.0013 ± 0.0006

[a]time since trigger

[b]flux in units of 10^{-11} erg cm^{-2} s^{-1} 2-10 keV

be added to the statistical error. The corrections and the systematic error are posted.[1] The count spectra are binned between 16 and 148.8 keV in ~2 keV bins. Table 1 summarizes the spectral fits to the entire burst (12.80 s) and to the peak 1 s, with 90% confidence limits. We fit the count spectra with three nested models, here presented as energy spectra[2]: a power law $F(E) \propto E^{\beta}$; a power law with an exponential cutoff $F(E) \propto E^{\beta} \exp[-E/E_0]$; and the 'Band' model (Band et al. 1993), a low energy power law with an exponential cutoff that transitions into a high energy power law $F(E) \propto E^{\beta_2}$. The peak energy $E_p = (1+\beta)E_0$ is both physically more relevant and less correlated with β than E_0; E_p is the energy of the peak of $E\ F(E) \propto \nu f_\nu$. The power law with an exponential cutoff is the same as the Band model with $\beta_2 = -\infty$, and the power law model is the same as the other two models with $E_0 = \infty$.

Table 1. BAT Spectral Fits

Parameter	Entire Burst			Peak Flux		
	Power Law[a]	Power Law, Cutoff[b]	Band Model[c]	Power Law[a]	Power Law, Cutoff[b]	Band Model[c]
β[d]	-0.78[e]	$0.01^{+0.11}_{-0.12}$	$0.01^{+0.11}_{-0.12}$	-0.42[e]	$0.45^{+0.14}_{-0.14}$	$-0.45^{+0.14}_{-0.14}$
β_2[f]	—	—	$-7.84^{+6.40}_{-1.16}$	—	—	$-8.27^{+7.14}_{-0.73}$
E_p[g]	—	$78.8^{+3.9}_{-3.1}$	$78.8^{+3.7}_{-3.1}$	—	$102.4^{+7.1}_{-6.3}$	$102.4^{+8.1}_{-6.3}$
Norm	7.15[e,h]	$14.2^{+5.9}_{-4.2}$ [i]	$15.0^{+1.65}_{-1.45}$ [h]	19.35[e,h]	$7.62^{+3.91}_{-2.65}$ [i]	$44.00^{+6.05}_{-5.20}$ [h]
χ^2/dof	181.7/57	15.2/56	15.2/55	169.8/57	31.6/56	31.6/55
p	$<.005$	$>.99$	$\geq .99$	$<.005$	$\geq .95$	$>.95$

[a]Power law model, $F(E) \propto E^{\beta}$.

[b]Power law with an exponential cutoff, $F(E) \propto E^{\beta} \exp[-E/E_0]$.

[c]Band model (Band et al. 1993), a low energy power law with an exponential cutoff transitioning to a high energy power law $F(E) \propto E^{\beta_2}$.

[d]The low energy spectral index.

[e]Fit too poor to produce uncertainty range.

[f]The high energy spectral index. The fit is insensitive to $\beta_2 < -2.5$ for the fitted E_p.

[g]The energy of the peak of $EN(E) \propto \nu f_\nu$, and $E_p = (2+\beta)E_0$.

[h]The normalization of the spectrum at 50 keV, in keV cm^{-2} s^{-1} keV^{-1}.

[i]The normalization of the spectrum at 1 keV, in keV cm^{-2} s^{-1} keV^{-1}.

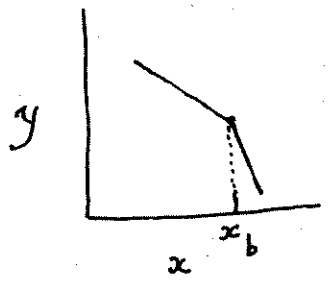
Number of free parameters

- A single polynomial of degree k

$$\alpha_0 + \alpha_1 x + \cdots + \alpha_k x^k$$

No. $= k + 1$

- Two lines and a break



$\alpha_1 + \beta_1 x$, $\alpha_2 + \beta_2 x$

(1, 2, 3, 4)

$\alpha_1 + \beta_1 x_b = \alpha_2 + \beta_2 x_b$ (restriction)

(5)

$5 - 1 = 4$

- Quadratic and line with a break

$\alpha_1 + \beta_1 x + \gamma_1 x^2$, $\alpha_2 + \beta_2 x + \gamma_2 x^2$, x_b

$6 - 1 = 5$

- Three lines and 2 breaks

$\alpha_1 + \beta_1 x$, $\alpha_2 + \beta_2 x$, $\alpha_3 + \beta_3 x$, x_{b1}, x_{b2} (breaks) or knots

$$\alpha_1 + \beta_1 x_{b1} = \alpha_2 + \beta_2 x_{b1}$$

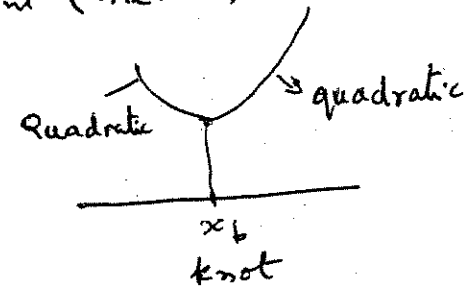
$$\alpha_2 + \beta_2 x_{b_2} = \alpha_3 + \beta_3 x_{b_2}$$

8 − 2 (restrictions) = 6 free parameters

- Two quadratics with a differentiable function at the break point (one knot)

$$\alpha_1 + \beta_1 x + \gamma_1 x^2 \qquad x < x_b$$

$$\alpha_2 + \beta_2 x + \gamma_2 x^2 \qquad x > x_b$$



$$\alpha_1 + \beta_1 x_b + \gamma_1 x_b^2 = \alpha_2 + \beta_2 x_b + \gamma_2 x_b^2$$

$$\beta_1 + 2\gamma_1 x_b = \beta_2 + 2\gamma_2 x_b$$

6 − 2 = 4 parameters

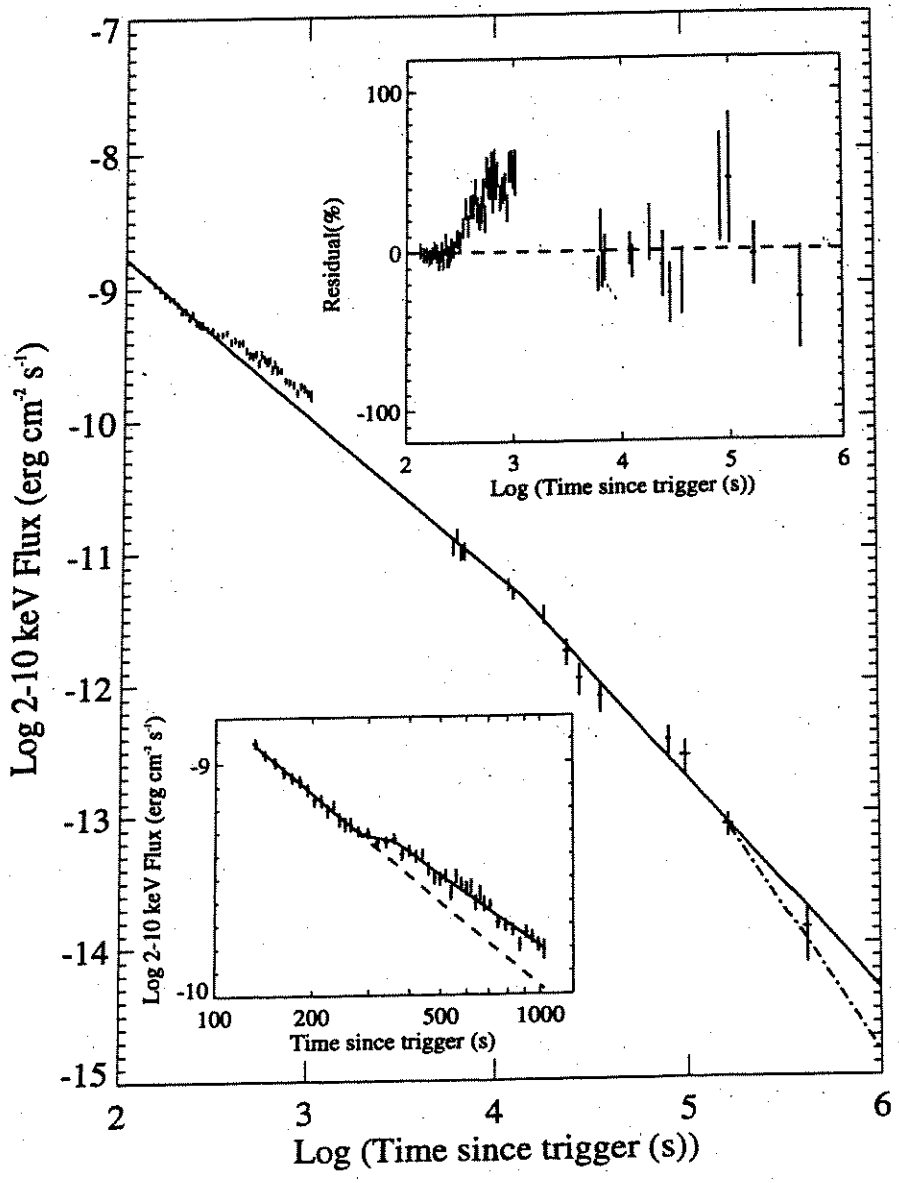


Fig. 3.— XRT decay light curve of GRB050525a including both Photodiode Mode (T<2000 s) and Photon Counting Mode (T>2000 s) data. The solid line is a broken power law fit to the combined data excluding those Photodiode Mode points colored green (see text). The dash-dot line is shown for illustration and has a slope of $\alpha = -2.2$, which is the value expected from simple modelling of a jet break (see text). The *lower inset* shows the data taken in photodiode mode only, during the first ∼1000 s after the BAT trigger. The solid line is a fit to the data with a power law model that includes two temporal breaks to different decay rates. The dashed line is an extrapolation of a simple power law fit (single slope) to the first segment of data prior to about 300 s. The *upper inset* shows the residuals with respect to the two power-law model fit to all the data, expressed as a percentage of the predicted model flux.

Table 3. UVOT multicolor data

T(mid)[a] (s)	T(exp) (s)	Mag	Flux Density[b]	T(mid)[a] (s)	T(exp) (s)	Mag	Flux Density[b]
V filter							
66	1.	13.21 ± 0.24	189.8 ± 41.2	148	5.	13.84 ± 0.12	106.2 ± 11.5
67	1.	12.90 ± 0.23	254.0 ± 53.6	153	5.	13.87 ± 0.12	103.2 ± 11.3
68	1.	12.86 ± 0.23	263.3 ± 55.5	158	5.	14.06 ± 0.12	87.1 ± 10.0
69	1.	13.01 ± 0.23	227.9 ± 48.5	163	5.	14.00 ± 0.12	91.7 ± 10.3
70	1.	12.97 ± 0.23	236.3 ± 50.1	168	5.	14.01 ± 0.12	90.8 ± 10.3
71	1.	13.31 ± 0.23	172.8 ± 34.0	173	10.	14.08 ± 0.13	83.3 ± 10.6
72	1.	13.13 ± 0.23	204.3 ± 43.9	258	10.	14.64 ± 0.14	49.8 ± 6.8
73	1.	13.01 ± 0.23	227.9 ± 48.5	342	10.	14.79 ± 0.15	43.3 ± 6.4
78	5.	13.13 ± 0.10	204.3 ± 19.6	426	10.	15.22 ± 0.17	29.2 ± 4.9
83	5.	13.26 ± 0.10	181.5 ± 17.7	511	10.	15.47 ± 0.19	23.2 ± 4.4
88	5.	13.18 ± 0.10	195.5 ± 18.9	595	10.	16.06 ± 0.24	13.4 ± 3.3
93	5.	13.24 ± 0.11	185.6 ± 18.1	680	10.	15.83 ± 0.22	16.6 ± 3.7
98	5.	13.25 ± 0.11	184.2 ± 17.9	764	10.	16.06 ± 0.25	13.4 ± 3.5
103	5.	13.51 ± 0.11	144.9 ± 14.7	849	10.	15.78 ± 0.22	17.4 ± 3.9
108	5.	13.44 ± 0.11	154.4 ± 15.5	933	10.	15.85 ± 0.24	16.3 ± 4.0
113	5.	13.67 ± 0.11	124.7 ± 13.0	1243	100.	16.34 ± 0.15	10.4 ± 1.5
118	5.	13.48 ± 0.11	148.4 ± 15.0	18575	156.	18.15 ± 0.41	2.0 ± 0.9
123	5.	13.62 ± 0.11	130.2 ± 13.5	22163	580.	19.10 ± 0.27	0.8 ± 0.2
128	5.	13.86 ± 0.12	104.2 ± 11.4	35638	750.	18.86 ± 0.27	1.0 ± 0.3
133	5.	13.70 ± 0.11	121.5 ± 12.8	49320	4982.	> 20.62	< 0.2
138	5.	13.83 ± 0.12	107.2 ± 11.6	971360	33800.	> 22.09	< 0.1
143	5.	13.81 ± 0.12	109.2 ± 11.8	1171176	6081.	> 21.16	< 0.1
B filter							
229	10.	14.79 ± 0.12	72.2 ± 8.4	904	10.	16.44 ± 0.20	15.8 ± 3.2
313	10.	15.19 ± 0.12	49.9 ± 5.8	1034	100.	16.61 ± 0.11	13.5 ± 1.4
397	10.	15.51 ± 0.13	37.2 ± 4.7	12671	390.	18.59 ± 0.18	2.2 ± 0.4
482	10.	15.63 ± 0.14	33.3 ± 4.6	16182	190.	18.69 ± 0.17	2.0 ± 0.3
571	10.	15.70 ± 0.14	31.2 ± 4.3	30031	388.	19.82 ± 0.52	0.7 ± 0.4
651	10.	16.13 ± 0.16	21.0 ± 3.3	33898	900.	20.84 ± 0.45	0.3 ± 0.1
735	10.	16.03 ± 0.16	23.0 ± 3.7	45468	896.	> 20.70	< 0.3
820	10.	16.56 ± 0.20	14.1 ± 2.9	62549	6513.	> 21.55	< 0.1
U filter							
215	10.	13.70 ± 0.18	110.3 ± 19.9	890	10.	15.29 ± 0.21	25.5 ± 5.4
299	10.	14.08 ± 0.18	77.8 ± 14.0	975	10.	15.32 ± 0.22	24.8 ± 5.6
419	10.	14.47 ± 0.19	54.3 ± 10.4	12019	900.	17.66 ± 0.17	2.9 ± 0.5

The background subtracted 2–10 keV light curve in the time interval T+128 s – T+1048 s (PD mode) is shown in Figure 3 (inset). The X–ray afterglow of GRB 050525 is clearly fading. The early afterglow decay was first fitted with a single power-law model, resulting in a best fit decay index $\alpha = -0.95 \pm 0.03$, with $\chi^2_r = 1.17$ (42 dof). Inspection of the residuals to the best fit model suggests that a flattening of the decay curve or a re-brightening of the source occurs at ~300 seconds after the trigger. A better fit is provided by a broken power law model with slopes α_1, α_2 and a break at t_b. This model gave $\chi^2_r = 0.98$ (40 dof), with best fit parameters $\alpha_1 = -1.23^{+0.03}_{-0.02}$, $\alpha_2 = -0.91$ and $t_b = 203$ s.

Again, however, the residuals suggest sytematic deviations from this model. We thus tried a broken power law with two temporal breaks. This model provided a very good fit to the data, with $\chi^2_r = 0.72$ (38 dof) and is plotted in Figure 3 (inset) as a solid line. The best fit parameters are $\alpha_1 = -1.19$, $t^1_b = 282$ s, $\alpha_2 = -0.30$, $t^2_b = 359$ s, and $\alpha_3 = -1.02$.

Next, we fitted the X–ray data taken in PC mode at times more than 5000 s after the trigger. We first used a single power-law model, obtaining a best fit decay index $\alpha = -1.51 \pm 0.07$, with $\chi^2_r = 1.40$ (12 dof). The poor fit is the result of a clear steepening of the light curve with time. We thus tried a broken power law model. The model provided a very good fit with $\chi^2_r = 0.97$ (10 dof) and best fit parameters $\alpha_1 = -1.16$, $\alpha_2 = -1.62$ and $t_b = 13177$ s.

Finally, we tried fitting the total light curve derived from the combined PD and PC mode data (see Figure 3). We find that the power law fit to the pre-brightening PD mode data (T<280 s) extrapolates well to the pre-break PC mode data. Moreover the decay index before 280 s agrees well with that of the PC mode data before the 13ks break. In contrast, if we extrapolate the post brightening PD mode data to later times using the best fit slope, a significant excess is predicted compared with the measured PC mode data. To join the post brightening PD mode data to the PC mode data requires a model with at least two temporal breaks, which are not constrained because of the intervening gap in X-ray coverage. We conclude that the brightening at about 280 s in the PD mode data represents a flare in the X-ray flux, possibly similar to the sometimes much larger flares that are seen at early times in other bursts (Burrows et al 2005b; Piro et al. 2005), and that the flux returns to the pre-flare decay curve prior to the start of our PC mode data.

We thus fit the combined PD and PC mode data excluding PD data at times t >T+288 s (green points in Figure 3). A broken power law model provided a good fit (solid line of Figure 3), with with $\chi^2_r = 0.50$ (25 dof) and best fit parameters $\alpha_1 = -1.20 \pm 0.03$, $\alpha_2 = -1.62^{+0.11}_{-0.16}$ and $t_b = 13726^{+7469}_{-5123}$ s. The break time is thus ~ 3.8 hours.

The complete XRT data are recorded in Table 2.

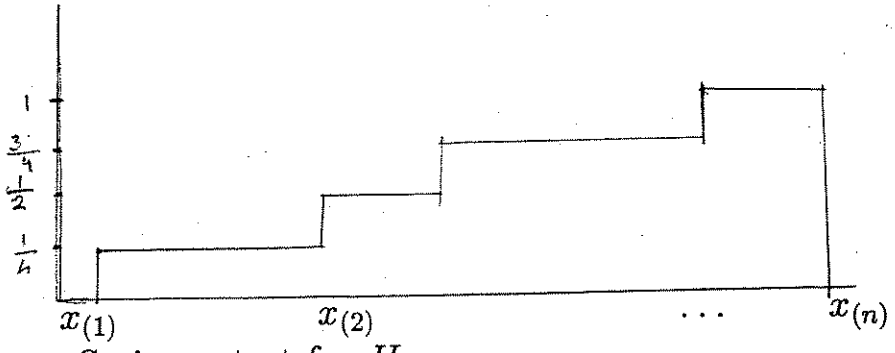
KOLMOGOROV-SMIRNOV AND CRAMÈR-VON MISES TESTS

Tests of simple hypothesis

$$H_{os} : F(x) \text{ is specified as } F_0(x)$$
$$H_1 : F(x) \text{ is arbitrary}$$

where $F(x)$ is distribution function of the random variable X.

Based on the sample $S = (x_1, \ldots, x_n)$ the ML estimate of $F(x)$ is the empirical distribution function, a step function, $\hat{F}_n(x)$ as shown in the graph. Each step is of magnitude $(1/n)$.



Kolmogorov-Smirnov test for H_0

$$KS_s = \sup_x \left| \hat{F}_n(x) - F_0(x) \right| .$$

Cramèr-von Mises test for H_0

$$CM_s = \int \left(\hat{F}_n(x) - F_0(x) \right)^2 dF_0(x).$$

The percentile points in each case can be obtained from bootstrap distribution.

TEST OF COMPOSITE HYPOTHESIS

$$H_{oc} : F(x) \in \{F(x,\theta), \theta \in \Theta\}$$
$$H_1 : F(x) \text{ is arbitrary.}$$

Let $\hat{\theta}$ be the MLE of θ based on the sample S. Then KS test statistic for H_{oc} is

$$KS_c = \sup_x \left| \hat{F}_n(x) - F(x,\hat{\theta}) \right| dF(x,\hat{\theta})$$

and CM test statistic for H_{oc} is

$$CM_c = \int \left(\hat{F}_n(x) - F(x,\hat{\theta})^2 dF(x,\hat{\theta}) \right)$$

The percentile points can be estimated by bootstrap sampling from $F(x,\hat{\theta})$.

References:

1. Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer.
2. Babu, G. and Rao, C.R. (2004) Goodness-of-fit tests when parameters are estimated. *Sankhya* 66, 63-74

INFORMATION THEORETIC CRITERIA FOR MODEL SELECTION

Problem

Let $g(x)$ be the true unknown probability distribution which gave rise to observed data, $S = (x_1, \ldots, x_n)$, a sample of n independent observations. Suppose that we have a set of r candidate models

$$f_i(x, \theta_i),\ \theta_i \in \Theta_i, \text{is a } k_i\text{-vector parameter},\ i = 1, \ldots, r.$$

The true $g(x)$ may or may not belong to any of the models.

The Kullback-Leibler measure of separation of the model f_i from g is

$$K(g, f_i) = \min_{\theta_i} \int \log \frac{g(x)}{f_i(x, \theta_i)} dG =$$

$$\int \log g(x) dG(x) - \max_{\theta_i} \int \log f(x, \theta_i) dG(x)$$

where G is the distribution function DF of g. The best choice of the model is s such that

$$\max_{\theta_s} \int \log f_s(x, \theta_s) dG \geq \max_{\theta_i} \int \log f_i(x, \theta_i) dG \forall i. \quad (C)$$

We call such $f_s(x, \theta_s)$ a quasi-true model. The criterion (C) cannot be used as G is unknown. However we have the empirical DF, $\hat{G}_n$ of G, based on the observed sample. Substituting $\hat{G}_n$ for G, we have

$$\max_{\theta_i} \int n \log f_i(x, \theta_i) d\hat{G} = \Sigma \log f_i(x, \hat{\theta}_i) = l_i(\hat{\theta}_i | S) \quad (L)$$

which is maximum likelihood under the model f_i

Unfortunately, the log likelihood in (L) is a biased estimate of the right hand side of (C), We estimate the bias and write the criterion as

$$C_i = -2l_i(\hat{\theta}_i|S) + 2b_i(\hat{G}_n) \qquad (D)$$

and choose the model for which C_i is a minimum. There are various choices of $b_i(\hat{G}_n)$, and the criteria based on (D) are called information theoretic criteria (ITC). Some choices which have been made in various applications are as follows.

1. $AIC = -2l_i(\hat{\theta}_i|S) + 2k_i$ (E)

where k_i is the number of parameters in model i, is called Akaike information criteria The choice of $b_i(\hat{G}_n)$ as k_i is strictly valid if g the true model, is a member of the i-th family.

2. Slight improvement over (E) are given in (F)

$$AIC_c = -2l_i(\hat{\theta}_i|S) + 2k_i[n/(n-k_i-1)] \qquad (F).$$

3. Takenchi information criterion

$$TIC = -2l_i(\hat{\theta}_i|S) + 2\,tr\left[J(\hat{\theta}_i)I(\theta_i)^{-1}\right] \qquad (G)$$

where

$$I(\hat{\theta}_i) = \left(I_{rs}(\hat{\theta}_i)\right)$$

$$I_{rs}(\hat{\theta}_i) = \left|\frac{\partial^2 l(\theta_i|S)}{\partial\theta_{ir}\partial\theta_{is}}\right|_{\hat{\theta}_i}$$

$$J(\hat{\theta}_i) = \Sigma_1^r a_r a_r'$$

$$a_r' = \left(\frac{\partial}{\partial\theta_{i1}}\log f_i(x_r|\theta_i), \ldots, \frac{\partial}{\partial\theta_{ik_i}}\log f_i(x_r|\theta_i)\right)|_{\hat{\theta}_i}.$$

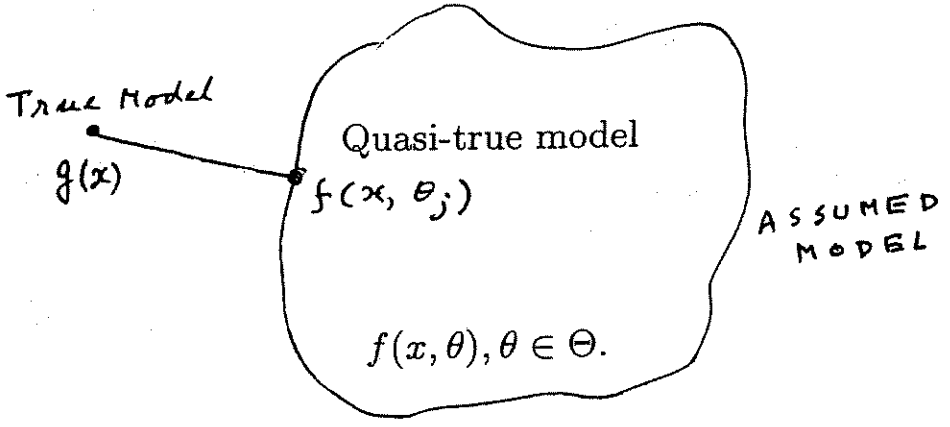
KULLBACK-LEIBLER (KL) MEASURE OF SEPARATION

KL is a measure of separation of a probability distribution $f(\theta, x)$ from a specified distribution $g(x)$.

$$KL(f,g) = \int f(x,\theta) \log \frac{f(x,\theta)}{g(n)} dx \geq 0$$

where

$$-\log \frac{f(x)}{g(x)} \text{ is Boltzman Entropy.}$$



$$\theta_g = \arg\min_{\theta} KL\left(f(x,\theta), g(x)\right).$$

Property of maximum likelihood

$$\hat{\theta} \to \theta_g \text{ as } n \to \infty.$$

Reference:

1. Nishi, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Analysis*, **27**, 392-403.

CRITERIA PROVIDING CONSISTENT ESTIMATE OF QUASI-TRUE MODEL

The information criteria based on bias corrected maximum log likelihood such as AIC, TIC, etc., may not provide consistent estimate of the quasi-true model, i.e., model closest to the true model in the set of candidate models as $n \to \infty$. Alternative criteria developed to provide consistent estimates as $n \to \infty$ are of the form

$$-2l_i\left(\hat{\theta}_i|S\right) + C_n k_i$$

where C_n is chosen such that

$$\lim_{n\to\infty} \frac{C_n}{n} = 0 \text{ and } \frac{C_n}{\log\log n} = +\infty.$$

Some choices of C_n are $\log n$, a $\log(\log n)$ with $a > 2$, and $\log n$. Bai, Rao and Wu (1999) presented a data oriented value for C_n.

Reference:

1. Bai, Z.D., Rao, C.R. and Wu, Y. (1999). Model selection with data oriented penalty, *J. Statist. Plann. Inference*, **77**, 103-117.

$BIC\ (\text{or } SBC) = -2\ell_i(\hat{\theta}_i|S) + k_i \log n$

SBC : Schwarz's Bayesian criterion provides an approximate Bayes factor

Table III. Akaike's information criterion (AIC) and Schwarz's Bayesian criterion (SBC) for six covariance structures.

	Structure name	AIC	SBC
1.	Simple	+459.5	+461.6
2.	Compound symmetric	+175.6	+179.9
3.	Autoregressive (1)	+139.5	+143.8
4.	Autogressive (1) with random effect for patients	+126.5	+132.9
5.	Toeplitz (banded)	+121.9	+139.2
6.	Unstructured	+110.1	+187.7

where $L(\hat{\theta})$ is the maximized log-likelihood or restricted log-likelihood (REML), q is the number of parameters in the covariance matrix, p is the number of fixed effect parameters and N^* is the total number of 'observations' (N for ML and $N - p$ for REML, where N is the number of subjects).

Models with ~~large~~ small AIC or SBC values indicate a better fit. However, it is important to note that the SBC criterion penalizes models more severely for the number of estimated parameters than does AIC. Hence the two criteria will not always agree on the choice of 'best' model. Since our objective is parsimonious modelling of the covariance structure, we will rely more on the SBC than the AIC criterion.

AIC and SBC values for the six covariance structures are shown in Table III. 'Unstructured', has the largest AIC, but AR(1)+RE, 'autoregressive with random effect for patient', has the largest SBC. Toeplitz ranks second in both AIC and SBC. The discrepancy between AIC and SBC for the UN structure reflects the penalty for the large number of parameters in the UN covariance matrix. Based on inspection of the correlation estimates in Tables I and III, the graphs of Figure 5, and the relative values of SBC, we conclude that AR(1)+RE, 'autoregressive with random effect for patient', is the best choice of covariance structure.

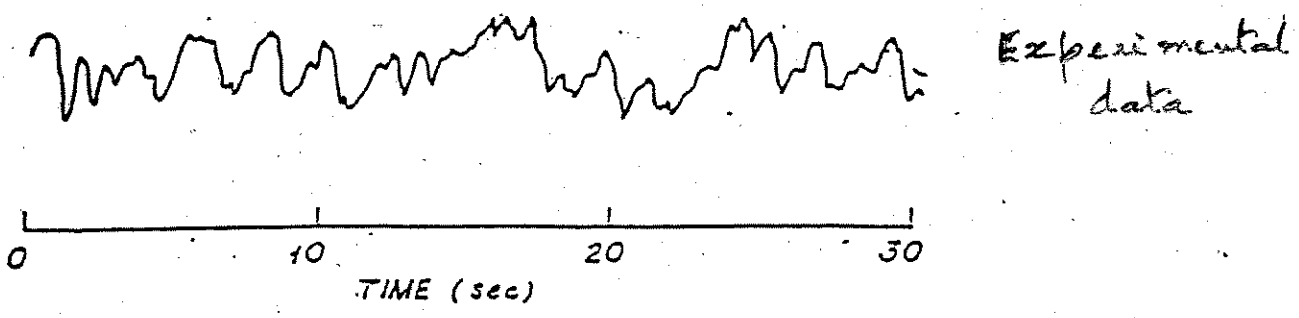
CROSS VALIDATION

Cross validation is generally used for variable selection in regression problems. Suppose that y is the variable to be predicted based on a set of k predictor variables, $x_1, \ldots, x_k$. Consider n sets of observations on $y, x_1, \ldots, x_k$.

y	x_1	x_2	$\ldots$	x_k
y_1	x_{11}	x_{21}	$\ldots$	x_{k1}
.	.	.	.	.
y_n	x_{1n}	x_{2n}	$\ldots$	x_{kn}

Divide the observations into two random sets of sizes n_1 and n_2 (usually with $n_1 << n_2$). Using the n_2 observations, which we may call the training set, we fit the regression or prediction functions $f_1(x), f_2(x), \ldots$ based on different subsets of variables. Use the different prediction functions to predict the y values in the n_1 observations not used, called the validating set, and for each function compute S=sum of squares of the differences between the observed and predicted values. Choose that function which gives the smallest value to S. The variables used to estimate the function are considered to be the best for predicting y. We can repeat the process a number of times by dividing the observations at random into training and validating sets and each time compute the sum of squares of differences for each subset. Finally the decision can be taken on the average of S values for each subset.

An interesting example due to the famous mathematician Mark Kac (see his autobiography *Enigmas of Chance*, pp. 74–76.) shows how the graph of a deterministic function could mimic the tracing of a random mechanism. To test Smoluchowski's theory of Brownian motion of a little mirror suspended on a quartz fiber in a vessel containing air, Kappler conducted an ingenious experiment in 1931 to obtain photographic tracings of the motion of the mirror. One such a tracing of 30 seconds' duration is reproduced in the figure below.



1

Kac remarks that looking at the graph, "it is difficult to escape the feeling that one is in the presence of chance incarnate and the tracing could only have been produced by a random mechanism." Kappler's experiment might be interpreted as confirming Smoluchowski's theory that the mirror is hit at random by the molecules of air giving the graph of the displacement of the mirror the character of a stationary Gaussian process.

Kac shows that the same kind of tracing indistinguishable from Kappler's graph by any statistical analysis, can be produced by plotting the function

$$\alpha \frac{\cos \lambda_1 t + \cos \lambda_2 t + \ldots + \cos \lambda_n t}{\sqrt{n}}$$

chaos

for sufficiently large n, choosing a sequence of numbers $\lambda_1, \ldots, \lambda_n$ and a scale factor α. Kac asks : So what is chance ?

modeling real data
= mostly deterministic + stochastic
(chaos)

7

If your experiment needs statistics, you ought to have done a better experiment.

- *Lord Rutherford*

A theory can be proved by an experiment but no path leads from experiment to theory.

- *Einstein*

It is safe to say that no discovery of some importance would have been missed by lack of statistical knowledge.

- *F.N. David*

There is no need for these hypotheses to be true, or even to be at all like the truth; rather one thing is sufficient for them –that they should yield calculations which agree with the observations.

Osiander
in preface to Copernicus *De Revolutionibus*

Osiander was a Protestant Theologian. The issue became more heated in the following century in the dispute between Galileo and the catholic church. The position of the latter as stated by Cardinal Bellarmino in 1615 was that the church would raise no objections if Galileo stated his theory as a mathematical hypothesis,"invented and assumed in order to abbreviate and ease the calculations", provided he did not claim it to be a true description of the world.

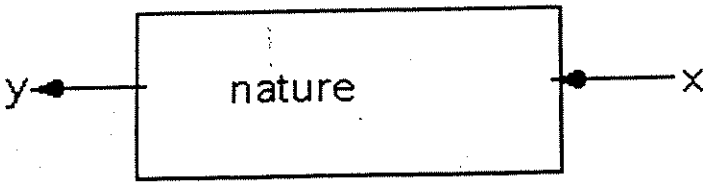
Statistics starts with data.

Think of the data as being generated by a black box .

A vector of input variables x (independent variables) go into one side.

Response variables y come out on the other side.

Inside the black box, nature functions to associate the input variables with the response variables, so the picture is like this:



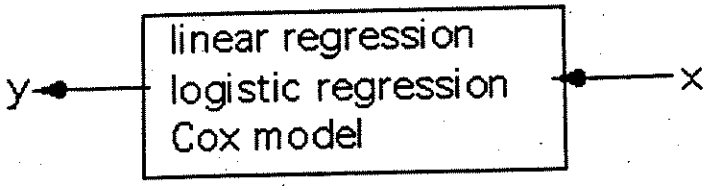
starts with assuming a stochastic data model for the inside of the black box.

A common data model is that data are generated by independent draws from:

response variables=f(predictor variables, random noise, parameters)

Parameters are estimated from the data and the model then used for information and/or prediction.

The black box is filled in like this:



Model Validation: yes-no using goodness-of-fit tests and residual examination

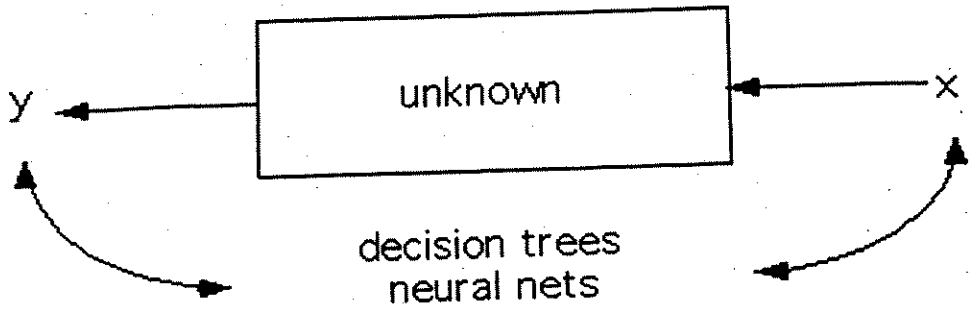
Estimated Culture Population: 98% of all statisticians

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown.

The approach is to find a function f(**x**)--an algorithm that operates on **x** to predict the responses **y**.

The black box looks like this



Model Validation: measured by predictive accuracy

Estimated Culture Population: 2% of statisticians--many in other fields

Two Major Goals In Analyzing The Data:

prediction: *to be able to predict what the responses are going to be to future input variables.*

regression:
classification:
accuracy:

Information *to extract some information about how nature is associating the response variables to the input variables.*

Classification methodologies

1. Support Vector Machines (SVM) (Vladimir Vapnik)
2. Classification trees (CART) (Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone)
3. Linear Discriminant Analysis (LDA) (Ronald Aylmer Fisher)
4. Quadratic Discriminant Analysis (QDA) (Ronald Aylmer Fisher)
5. Neural Networks (Nnet)

6. Generalized Linear Model (GLM) (Example: Logistic Regression Model)
7. Multinomial Logit Models (MLM) (Classification into more than 2 categories)
8. Nearest Neighbor (NN)
9. Learning Vector Quantization (LVQ)
10. Flexible Discriminant Analysis (FDA)
11. Mixture Discriminant Analylsis (MDA)
12. Nonparametric Density Estimation and Likelihood

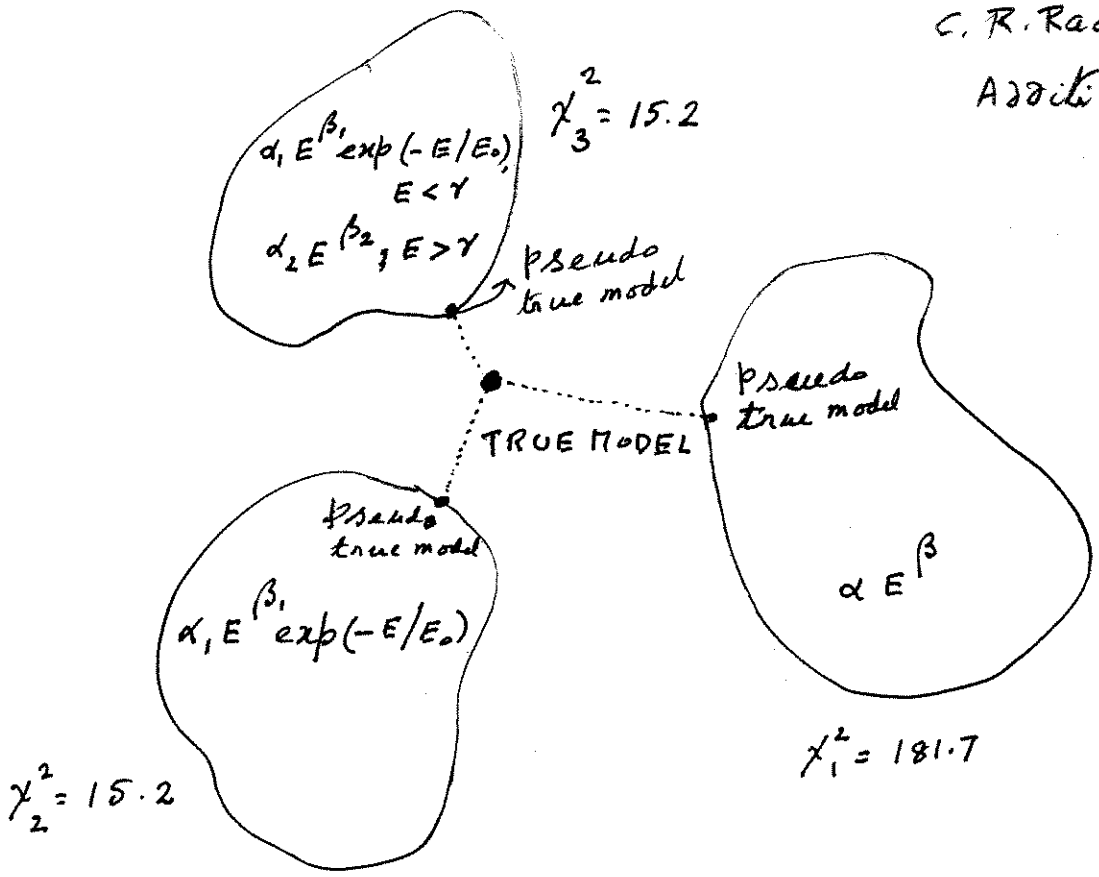
Reference

ANALYSIS OF MICROARRY GENE EXPRESSION DATA

3 - Mei-Ling Ting Lee (Kluwer Academic Publishers)

C. R. Rao

Additional notes



MODEL SELECTION CRITERIA (MSC)

- Choose the model
- Do Cross validation for accuracy of Prediction
- If MSC suggests more than one model we may have to find some sort of average model!

It would be interesting to apply these model choosing methods on the gamma ray burst data

Check χ_2^2 and χ_3^2 !!

Information matrix

$$I(\theta) = \left(i_{rs} \right)$$

$q \times q$ matrix $\qquad$ $q \times q$

$$i_{rs} = E(S_i S_j) \qquad S_i = i\text{-th score function}$$

Important result

$$\hat{\theta} \sim \text{(asymptotically)} \; N_q(\theta, I^{-1}) \quad \text{as } n \to \infty$$

under some conditions.