NOTES ON

MODEL SELECTION AND EVALUATION

GOODNESS-OF-FIT AND LIKELIHOOD RATIO TESTS

C. R. Rao

Summer School in statistics for astronomers

June 7, 2005

DATA

- PROBLEM ORIENTED

  - Testing a proposed theory

Examples

Cosmic magnification of the universe as predicted by the General Theory of Relativity (Einstein)

Are quasars local or distant?

Fred Hoyle's theory of steady state of the Universe ($N = S^{-1.5}$)

- Appropriate data to be collected by

DESIGN OF EXPERIMENTS

SAMPLE SURVEY

- EVENT ORIENTED

Take measurements when an event of interest occurs.

Gama Ray Bursts

- TRANSACTIONAL DATA

Records of Insurance claims, Sales in a store...

STATISTICAL INFERENCE

- MODEL BASED ANALYSIS (Specification) R. A. Fisher

  - $(\mathcal{X}, \mathcal{B}, \mathcal{P}_\theta)$ Kolmogorov set up

    $\mathcal{P}_\theta$ = class of probability distributions of which generates the observed data

    $\mathcal{X}$ = space of observed data

  - Prior information on θ

    Bayes and empirical Bayes

- NONPARAMETRIC METHODS

  Randomization and Rank tests

- BLACK BOX APPROACH

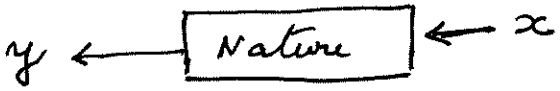
  Algorithmic

  Neural Networks, CART, ...

Statistical Modeling: The two cultures

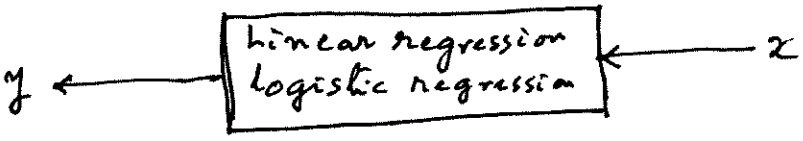
(Statistical Science, 2001, 16, 199-231

by Leo Breiman

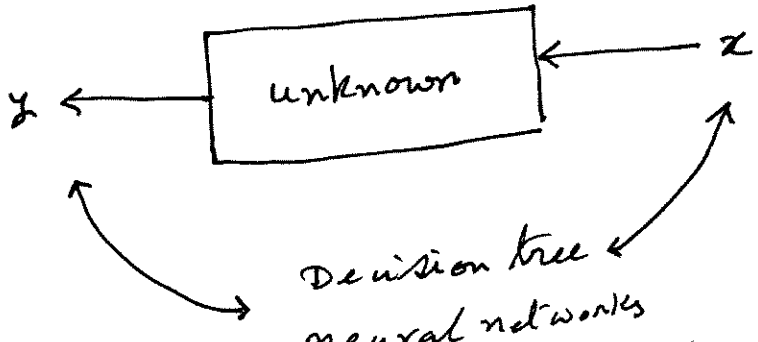


Data modeling culture (98% of statisticians)

Assume (given) Stochastic model



Algorithm modeling culture (2% of statisticians)



Model Validation: Predictive accuracy by Cross Validation

METHODOLOGICAL PROBLEMS OF STATISTICS

Fisher's formulation of problems in statistics

- Specification (Prob. model for data)
- Estimation (choosing a model from a given class)
- Distribution (Tests of hypotheses)

(Fisher 1922, Borel 1909, Pearson 1902)

"As regards problems of specification, these are entirely a matter for the practical statistician, for these cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type. In other cases we may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested a *posteriori*."

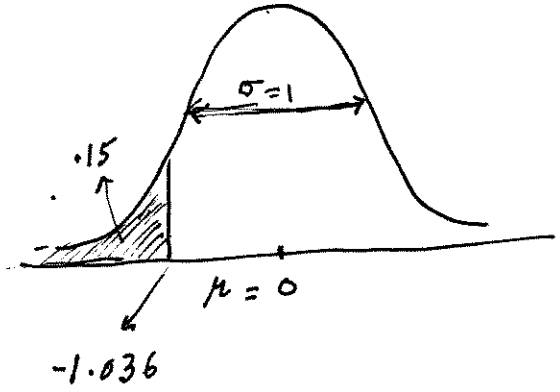
TEST FOR NORMALITY

(Constructing a Q–Q plot)

A sample of $n = 10$ observations gives the values in the following table:

Ordered observations $x_{(j)}$	Probability levels $(j - \frac{1}{2})/n$	Standard normal quantiles $q_{(j)}$
−1.00	.05	−1.645
−.10	.15	−1.036
.16	.25	−.674
.41	.35	−.385
.62	.45	−.125
.80	.55	.125
1.26	.65	.385
1.54	.75	.674
1.71	.85	1.036
2.30	.95	1.645

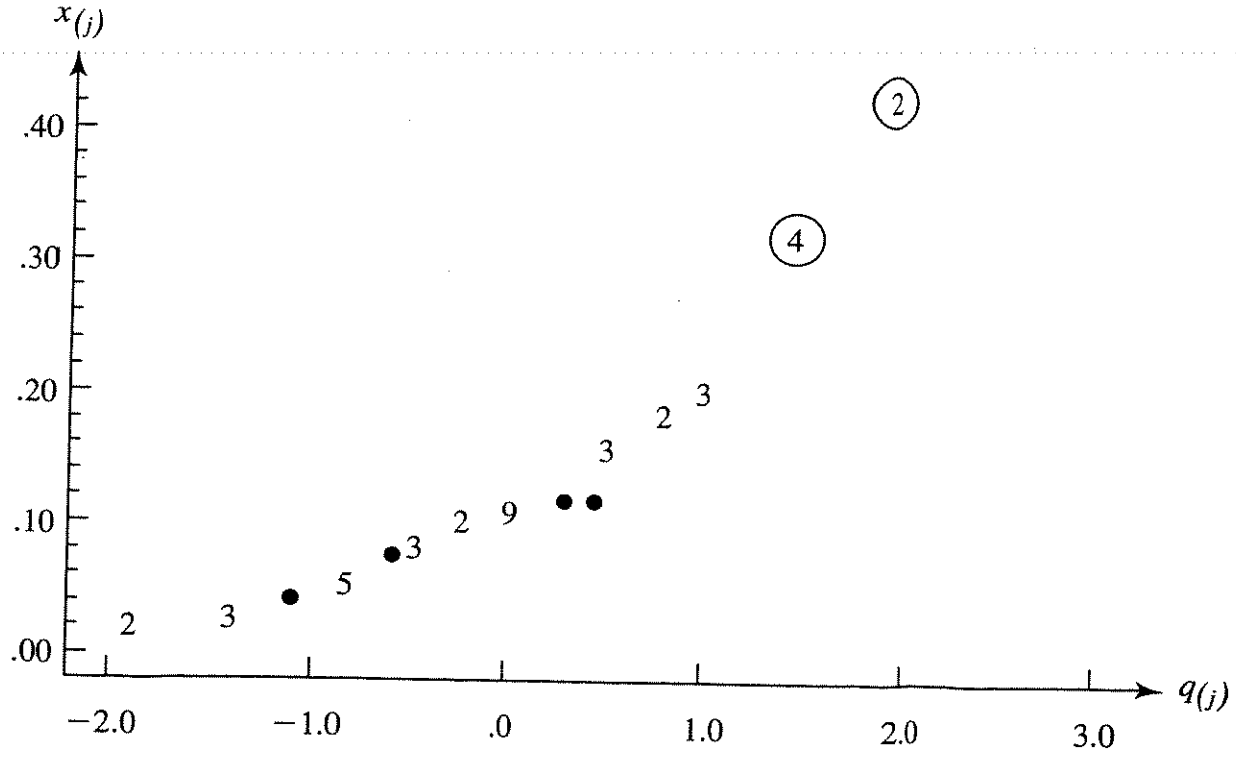
Here, for example, $P[Z \leq .385] = \int_{-\infty}^{.385} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = .65.$



4.1 RADIATION DATA (DOOR CLOSED)

Radiation	Oven no.	Radiation	Oven no.	Radiation
.15	16	.10	31	.10
.09	17	.02	32	.20
.18	18	.10	33	.11
.10	19	.01	34	.30
.05	20	.40	35	.02
.12	21	.10	36	.20
.08	22	.05	37	.20
.05	23	.03	38	.30
.08	24	.05	39	.30
.10	25	.15	40	.40
.07	26	.10	41	.30
.02	27	.15	42	.05
.01	28	.09		
.10	29	.08		
.10	30	.18		

ce: Data courtesy of J. D. Cryer.



A Q–Q plot of the radiation data (door closed) from Example 4.10. (The integers in the plot indicate the number of points occupying the same location.)

$$r_Q = \frac{\sum_{j=1}^{n} (x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n} (x_{(j)} - \bar{x})^2} \sqrt{\sum_{j=1}^{n} (q_{(j)} - \bar{q})^2}}$$

Formally, we reject the hypothesis of normality at level of significance α if r_Q falls *below* the appropriate value in Table 4.2.

TABLE 4.2 CRITICAL POINTS FOR THE *Q–Q* PLOT CORRELATION COEFFICIENT TEST FOR NORMALITY

Sample size n	Significance levels α		
	.01	.05	.10
5	.8299	.8788	.9032
10	.8801	.9198	.9351
15	.9126	.9389	.9503
20	.9269	.9508	.9604
25	.9410	.9591	.9665
30	.9479	.9652	.9715
35	.9538	.9682	.9740
40	.9599	.9726	.9771
45	.9632	.9749	.9792
50	.9671	.9768	.9809
55	.9695	.9787	.9822
60	.9720	.9801	.9836
75	.9771	.9838	.9866
100	.9822	.9873	.9895
150	.9879	.9913	.9928
200	.9905	.9931	.9942
300	.9935	.9953	.9960

Box and Cox [3] consider the slightly modified family of power transformations

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases} \tag{4–34}$$

which is continuous in λ for $x > 0$. (See [8].) Given the observations $x_1, x_2, \ldots, x_n$, the Box–Cox solution for the choice of an appropriate power λ is the solution that *maximizes* the expression

$$\ell(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} (x_j^{(\lambda)} - \overline{x^{(\lambda)}})^2 \right] + (\lambda - 1) \sum_{j=1}^{n} \ln x_j \tag{4–35}$$

We note that $x_j^{(\lambda)}$ is defined in (4-34) and

$$\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{x_j^{\lambda} - 1}{\lambda} \right) \tag{4–36}$$

is the arithmetic average of the transformed observations. The first term in (4–35) is, apart from a constant, the logarithm of a normal likelihood function, after maximizing it with respect to the population mean and variance parameters.

The calculation of $\ell(\lambda)$ for many values of λ is an easy task for a computer. It is helpful to have a graph of $\ell(\lambda)$ versus λ, as well as a tabular display of the pairs $(\lambda, \ell(\lambda))$, in order to study the behavior near the maximizing value $\hat{\lambda}$. For instance, if either $\lambda = 0$ (logarithm) or $\lambda = \frac{1}{2}$ (square root) is near $\hat{\lambda}$, one of these may be preferred because of its simplicity.

Rather than program the calculation of (4-35), some statisticians recommend the equivalent procedure of fixing λ, creating the new variable

$$y_j^{(\lambda)} = \frac{x_j^{\lambda} - 1}{\lambda \left[\left(\prod_{i=1}^{n} x_i \right)^{1/n} \right]^{\lambda - 1}} \qquad j = 1, \ldots, n \tag{4–37}$$

and then calculating the sample variance. The minimum of the variance occurs at the same λ that maximizes (4-35).

Comment. It is now understood that the transformation obtained by maximizing $\ell(\lambda)$ usually improves the approximation to normality. However, there is no guarantee that even the best choice of λ will produce a transformed set of values that adequately conform to a normal distribution. The outcomes produced by a transformation selected according to (4-35) should always be carefully examined for possible violations of the tentative assumption of normality. This warning applies with equal force to transformations selected by any other technique.

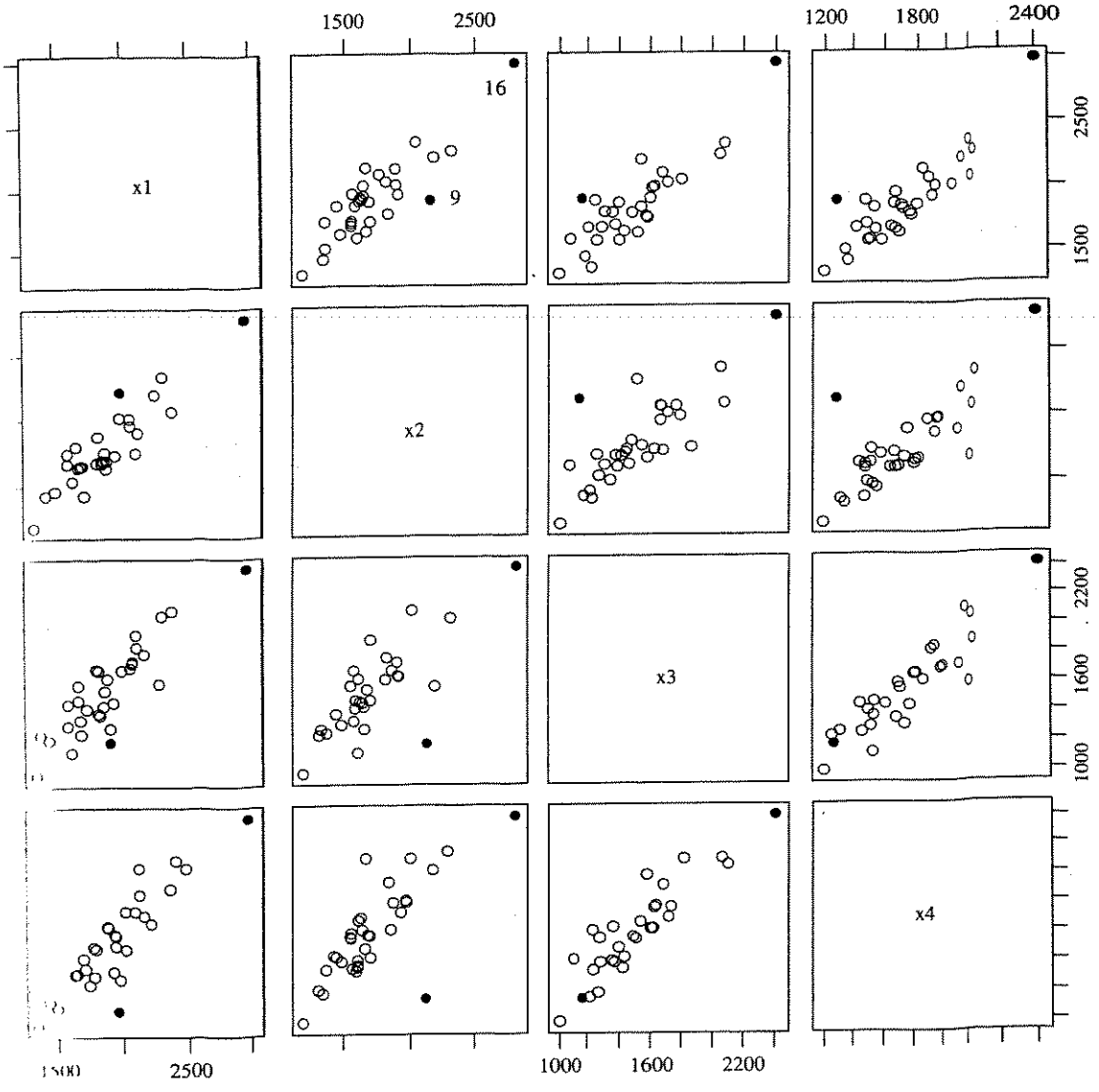
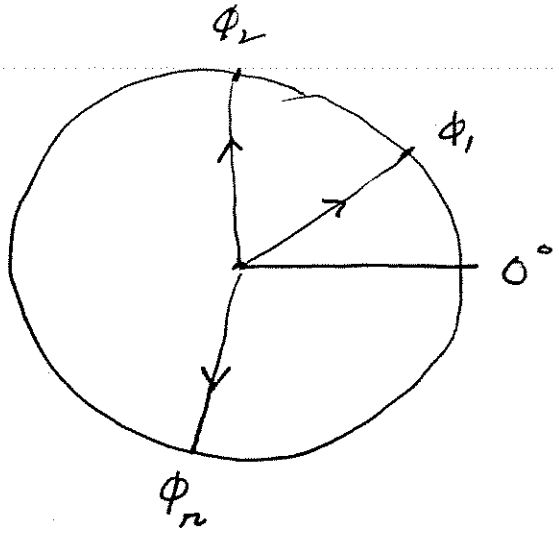


Figure 4.11 Scatter plots for the lumber stiffness data with specimens 9 and 16 plotted as solid dots.

TESTS OF RANDOMNESS ON A CIRCLE AND SPHERE

Circular (directional data on a circle)



Compute

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} \cos\phi_i \,, \quad \bar{y} = \frac{1}{n}\sum_{1}^{n} \sin\phi_i$$

Resultant vector $\underline{r}$ has length r

$$r = \left(\bar{x}^2 + \bar{y}^2\right)^{1/2}$$

1. Rayleigh Test: Reject H_0 of randomness if $r \geq r_\alpha$ or $z = nr^2 \geq z_\alpha$. For critical values see the book: Circular Statistics in Biology by Edward Batschelet, Academic Press

2. J. S. Rao Spacing Test

Compute

$$T_1 = \phi_2 - \phi_1, \; T_2 = \phi_3 - \phi_2, \cdots, T_{n-1} = \phi_n - \phi_{n-1}$$

$$T_n = 360 - (\phi_n - \phi_1)$$

Check $\Sigma T_i = 360$

$$U = \frac{1}{2} \sum_{1}^{n} \left| T_i - \frac{360}{n} \right|$$

Reject hypothesis of randomness if

$$U \geq U_\alpha$$

For critical values, see the Table in the book by Batschelet.

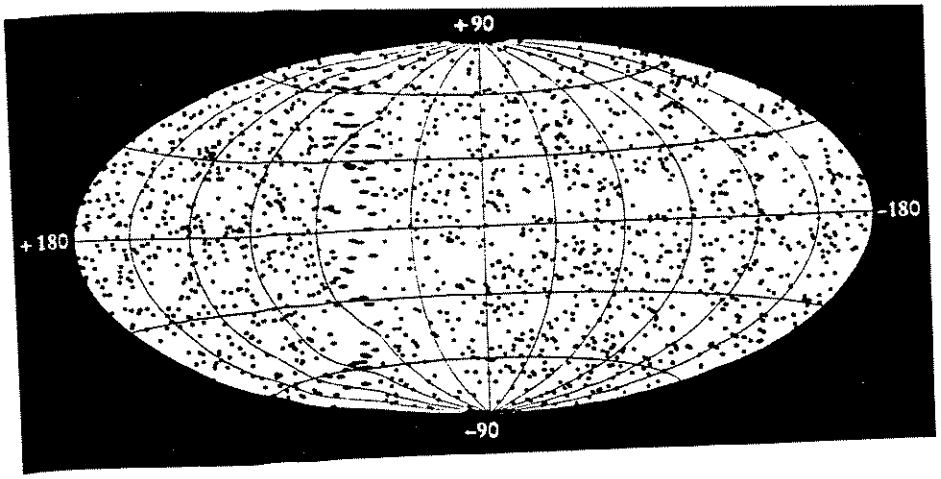


FIGURE 3. BATSE MAP IN GALACTIC COORDINATES of the 1636 bursts detected between April 1991 and August 1996. Each dot corresponds to a single gamma-ray burst. The distribution is consistent with an isotropic burst population. (Map courtesy of BATSE team.)

precision) by analyzing the images from the x-ray telescopes. Counterparts in the optical and radio wavebands can then be sought. Of the eight bursts examined so far, three have definite optical counterparts, one of which also has a radio counterpart.

In this rapidly unfolding story of discovery, most news circulates through International Astronomical Union (IAU) circulars and e-mail. At this point, the preliminary results are as follows.

▷ The first confidently identified optical counterpart was for GRB 970228, which was observed with a ground-based telescope on La Palma, one of the Canary Islands. A faint optical source (21st magnitude in the visible) was found to fade over a period of a week. Observations of the same position by the Hubble Space Telescope found diffuse emission surrounding the counterpart that could be a distant galaxy. No spectra were obtained, however.

▷ Another burst, GRB 970508, also had a weak, fading optical counterpart. For that burst, spectral lines were detected by the Keck II telescope in Hawaii with a redshift of 0.8. If these are lines from gas in a galaxy where the burst source resides, they imply a distance of 1.8 Gpc to the burster.

The new evidence strongly supports a cosmological origin for the bursts.

Active galactic nuclei

It is now widely accepted that the highly luminous (10^{40}–10^{46} ergs^{-1}) emission from the central cores of quasars and other active galactic nuclei (AGN) is produced by the accretion of gas onto a massive (10^6–10^9 $M_\odot$) black hole. The origins of this idea can be traced to suggestions made in the early 1960s by Fred Hoyle and William Fowler[6] and Edwin Salpeter[7] of supermassive objects powering quasars and to the work in the late 1960s and early 1970s of Donald Lynden-Bell[8] and Martin Rees,[9] who developed the framework of the current AGN paradigm. Results from Granat and Compton give additional support to this picture.

The new gamma-ray data show that there are two distinct classes of AGN defined by their redshift and luminosity distributions and their gamma-ray spectral properties. Sources in the first class, which are generally associated with AGNs classified in other wavelength ranges as Seyfert galaxies, have redshifts of less than 0.06 and 50–150 keV luminosities in the range of 10^{41}–10^{44} ergs^{-1}. As first discovered by Sigma in the case of NGC 4151, these sources display spectral steepening at energies around 60 keV. Their gamma-ray emission is thought to come from the inner accretion disk around the black hole, and the spectra are quite similar in shape to those of Galactic black holes such as Cyg X-1.

The second class of gamma-ray AGN consists of those associated with blazars—quasars with strong radio emission and flat radio spectra. With redshifts as large as 2.3, these objects have gamma-ray luminosities—assuming isotropic emission—as high as 10^{49} ergs^{-1}. In fact, the gamma-ray luminosity of these objects from 20 MeV to 30 GeV is often much higher than their luminosity at other wavelengths.

The discovery of the gamma-ray blazars was made by EGRET in 1991. To date, more than 60 gamma-ray blazars have been identified. Almost all of them vary strongly and flare for periods of days to weeks. A full-sky map of the gamma-ray sky as viewed by EGRET is shown in figure 4. The sources seen off the Galactic plane are the blazars.

Blazars are thought to be AGNs whose orientations are such that we observe them nearly along the axis of relativistic particle jets that emanate from a central black hole. The emission is strongly beamed, which implies that we see only a small proportion (10%) of all systems, and that the actual luminosities are significantly lower than the 10^{49} ergs^{-1} quoted above.

The future

The rapid development of gamma-ray astronomy that has occurred since the launches of Granat and Compton is likely to continue into the future as these missions observe more and as other approved and planned missions get under way. Granat is nearing the end of its life, but will be able to make a few more observations toward the center of our Galaxy over the coming years. Compton's onboard thrusters have recently been fired to raise its orbit enough to keep it in space until at least 2005. All four of the Compton instruments are working and have indefinite lifetimes, except for EGRET, which will run out of gas for its spark chamber in about two years.

Beyond Compton and Granat, there are two major missions being planned. The first is the International Gamma Ray Astrophysics Laboratory—INTEGRAL for short—which is an approved mission of the European Space Agency (ESA) with the participation of Russia and the US.[11] Its launch is scheduled for 2001. INTEGRAL's selected payload consists of two main instruments—an imager and a spectrometer—both of which are coded-aperture telescopes similar to Sigma but with improved detector technology.

Making use of cadmium telluride semiconductor detectors and cesium iodide scintillation detectors, INTE-

12

Distribution on a sphere

The data are in the form of unit vectors (3 dimensional)

$$l_{11} \quad l_{21} \quad l_{31}$$

$$\vdots$$

$$l_{1n} \quad l_{2n} \quad l_{3n}$$

The resultant vector

$$\bar{l}_1 \quad \bar{l}_2 \quad \bar{l}_3, \quad \bar{l}_i = \frac{1}{n} \Sigma$$

with ~~crossed~~ length R, where

$$R^2 = \bar{l}_1^2 + \bar{l}_2^2 + \bar{l}_3^2$$

Test: H_0 of randomness is rejected if

$$\frac{3R^2}{n} \geq \chi^2_\alpha \text{ (on 3 degrees of freedom)}$$

in large samples

TESTING GOODNESS OF FIT

- KARL PEARSON χ^2 TEST

Multinomial Distribution in k classes

classes	1	. . .	k	Total
Prob	$\pi_1(\theta)$	. . . .	$\pi_k(\theta)$	1
frequency	f_1	. . .	f_k	n

$\underset{\sim}{\theta}$ = p-vector parameter

Test H_0 : $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$ Specified (Simple hypothesis)

H_1 : Arbitrary

$$\chi^2 = \sum_1^k \frac{[f_i - n\pi_i(\theta_0)]^2}{n\pi_i(\theta_0)}$$, k−1 degrees of freedom (d.f.)

Composite Hypothesis

H_0 : $\theta \in (H)$

H_1 : Arbitrary

$$\chi^2 = \sum_1^k \frac{[f_i - n\pi_i(\hat{\theta})]^2}{n\pi_i(\hat{\theta})}$$, k − p − 1 d.f.

where $\hat{\theta}$ is obtained by minimizing $\sum^k [f_i - n\pi_i(\theta)]^2 / n\pi_i(\theta)$

• Continuous Distributions

$x_1, x_2, \ldots, x_n$ are n i.i.d observations from a prob. distribution G with density g (unknown).

Simple hypothesis

H_0 : $G(x) = H(x)$, $[g(x) = h(x)]$ (specified)

H_0 : $G(x) \neq H(x)$

1. χ^2 - test

Choose $a_0 = -\infty < a_1 < \cdots < a_k < \infty = a_{k+1}$

Let f_i = no. of x_i in the interval (a_i, a_{i+1})

$$m_i = H(a_{i+1}) - H(a_i)$$

$$\chi^2 = \sum_{i=0}^{k} \frac{[f_i - n m_i]^2}{n m_i}, \quad k \text{ d.f.}$$

2. Graphical method (Simple Hypotheses)

Let $x_{(1)}, \ldots, x_{(n)}$ be order statistics

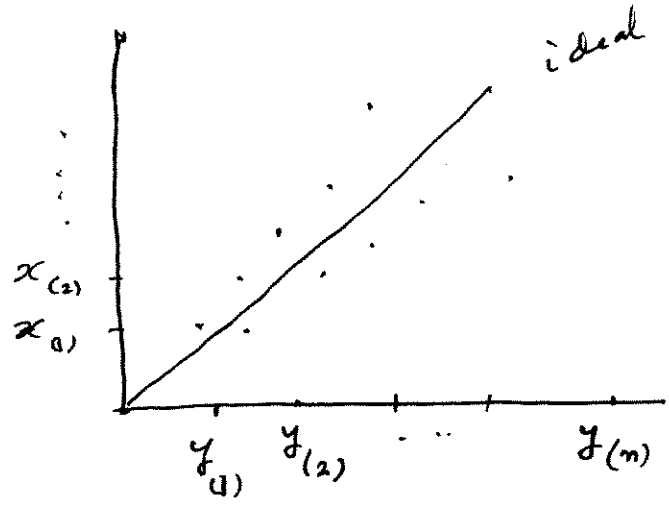
$$y_{(1)}, \ldots, y_{(n)}$$

where

$$\frac{2i-1}{2n} = \int_{-\infty}^{y_{(i)}} h(z)\, dz = H(y_{(i)}) - H(-\infty)$$

and $h(z)$ is the specified density.

Plot the curve



Test by resampling (bootstrap)

$$r^2 = \frac{\left[\Sigma (x_{(i)} - \bar{x})(y_{(i)} - \bar{y})\right]^2}{\Sigma (x_i - \bar{x})^2 \, \Sigma (y_{(i)} - \bar{y})^2}$$

3. Kolmogorov statistic (simple hypothesis)

$$K = \sup_x |G_n(x) - H(x)|$$

where $G_n(x)$ is the empirical distribution function based on the sample $x_1, \ldots, x_n$

$$G_n(x) = \frac{1}{n} \text{ (no. of observations} \le x)$$

$H(x)$ = Specified distribution function

Reject H_0 if

$$K \ge c$$

c is chosen such that

$$\text{Prob}(K \ge c) = \alpha \text{ (level of significance)}$$

A table of c values for $n \le 100$ is given by Owen (1962): Handbook of Statistical Tables, Addison-Wesley

For larger n use the formula

$$P(\sqrt{n}\, K \le z) \simeq 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 z^2}$$

CONTINUOUS DISTRIBUTIONS

Composite hypothesis

$$H_0: \quad G(x) = H(x, \theta), \quad \theta \in \Theta \subset R^p$$

$$H_1: \quad G(x) = \text{arbitrary}$$

1. χ^2 - test

Choose $a_0 = -\infty < a_1 < \cdots < a_k < \infty = a_{k+1}$

Let $f_0, f_1, \ldots, f_k$ be the observed frequencies in the $k+1$ intervals. Define

$$m_i(\theta) = H(a_{i+1}, \theta) - H(a_i, \theta)$$

Then

$$\chi^2(\theta) = \sum_{i=0}^{k} \frac{[f_i - n m_i(\theta)]^2}{n m_i(\theta)}$$

Test criterion

$$\chi^2(\hat{\theta}) = \min_{\theta} \chi^2(\theta)$$

is distributed on $k-p$ d.f., where p is the number of parameters and the no. of intervals is $k+1$

2. Graphical Method

Specified model $H(x, \theta)$

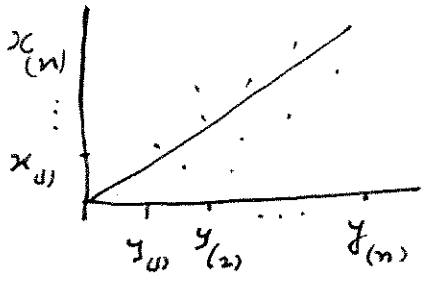
Ordered data $x_{(1)}, \cdots, x_{(n)}$

Let $\hat{\theta}$ be m.l. estimate of θ using observed data, $x_1, \cdots, x_n$

Compute $y_{(i)}$, $i = 1, \cdots, n$ such that

$$\frac{2i-1}{n} = H\left(y_{(i)}, \hat{\theta}\right) - H(-\infty, \hat{\theta})$$

Plot the curve



Compute

$$r^2 = \frac{[\Sigma(x_{(i)} - \bar{x})(y_{(i)} - \bar{y})]^2}{\Sigma(x_{(i)} - \bar{x})^2 \Sigma(y_{(i)} - \bar{y})^2}$$

Test $r^2 \leq r_\alpha^2$: Reject

The distribution of r^2 in samples from $H(x, \theta)$ may depend on θ (unknown). In some cases, for example if $H(x, \theta)$ belongs to the location-scale family, the distribution of r^2 does not depend on θ. One suggestion is to use the bootstrap distribution of r^2 drawing samples from $H(x, \hat{\theta})$. A hard way is to find $r_\alpha^2(\theta)$ drawing samples from $H(x, \theta)$ and choosing $\max r_\alpha^2(\theta)$ over θ as the critical value.

3. Kolmogorov-Smirnoff test

$$KS = \sup_x | G_n(x) - H(x, \hat{\theta}) |$$

The distribution of KS is complicated

Bootstrap critical values can be obtained by sampling from $H(x, \hat{\theta})$

The test will be conservative.

HOLY TRINITY

- LIKELIHOOD RATIO TEST

  NEYMAN-PEARSON (1928, 1933)

- WALD (1943) TEST

- RAO SCORE (1948) TEST

They provide tests of goodness of fit in the following situations.

H_0: $p(x, \theta)$, $\theta \in \Theta_1$

H_1: $p(x, \theta)$, $\theta \in \Theta \supset \Theta_1$ (nested case)

and some regularity conditions* are satisfied.

Reference

C. R. Rao (1973). Linear Statistical Inference and its Applications. Wiley

* For instance, these tests are not applicable for testing that the prob. density is a mixture of k normal distributions against the alternative

Let $x_1, \ldots, x_n$ be iid observations from $p(x, \underset{\sim}{\theta})$, $\underset{\sim}{\theta}$ a p-vector parameter. Define

$$l(\theta) = \sum_1^n \log p(x_i, \theta) \quad \text{(log likelihood)}$$

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

p-score vector

$$I(\theta) = -E\left[\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right] \quad p \times p \text{ information matrix}$$

H_0 may be written in the form:

$$\underset{\sim}{h}(\theta) = (h_1(\theta), \ldots, h_r(\theta))$$

$$\underset{\sim}{h}(\theta) = \underset{\sim}{c} : (h_1(\theta) - c_1 = 0, \ldots, h_r(\theta) - c_r = 0)$$

r dimensional hypothesis

$\underset{p\text{-vector}}{\hat{\theta}} = \arg\max_\theta l(\theta)$, unrestricted maximum likelihood estimator

$\hat{\theta}_r = \arg\max_{h(\theta) = \underset{\sim}{c}} l(\theta)$ restricted m.l. estimator

Likelihood Ratio Statistic

$$LR = 2\left[l(\hat{\theta}) - l(\hat{\theta}_n) \right]$$

Wald statistic

$$W = (\underset{\sim}{h}(\hat{\theta}) - \underset{\sim}{c})' \left[H(\hat{\theta})' I(\hat{\theta})^{-1} H(\hat{\theta}) \right]^{-1} (\underset{\sim}{h}(\hat{\theta}) - \underset{\sim}{c})$$

$$\underset{p \times r}{H(\theta)} = \begin{pmatrix} \frac{\partial h_1}{\partial \theta_1} & \cdots & \frac{\partial h_r}{\partial \theta_1} \\ \vdots & \cdots & \vdots \\ \frac{\partial h_1}{\partial \theta_p} & \cdots & \frac{\partial h_r}{\partial \theta_p} \end{pmatrix}$$

Rao's Score statistic

$$RS = s(\hat{\theta}_n)' I(\hat{\theta}_n)^{-1} s(\hat{\theta}_n)$$

$$s(\hat{\theta}_n) = (s_1(\hat{\theta}_n), \cdots, s_p(\hat{\theta}_n))$$

$$s_i(\hat{\theta}_n) = \frac{\partial l}{\partial \theta_i}\Big|_{\theta = \hat{\theta}_n}$$

All the three statistic have the asymptotic distribution as (chi-square) χ^2 on r d.f.

MODEL SELECTION

Problem: Given possible models for given data

$P(x, \theta_1), \ldots\ldots P(x, \theta_k)$

choose one as a parent for given data $(x_1, x_2, \text{---} x_n)$

Purpose: Predicting future events

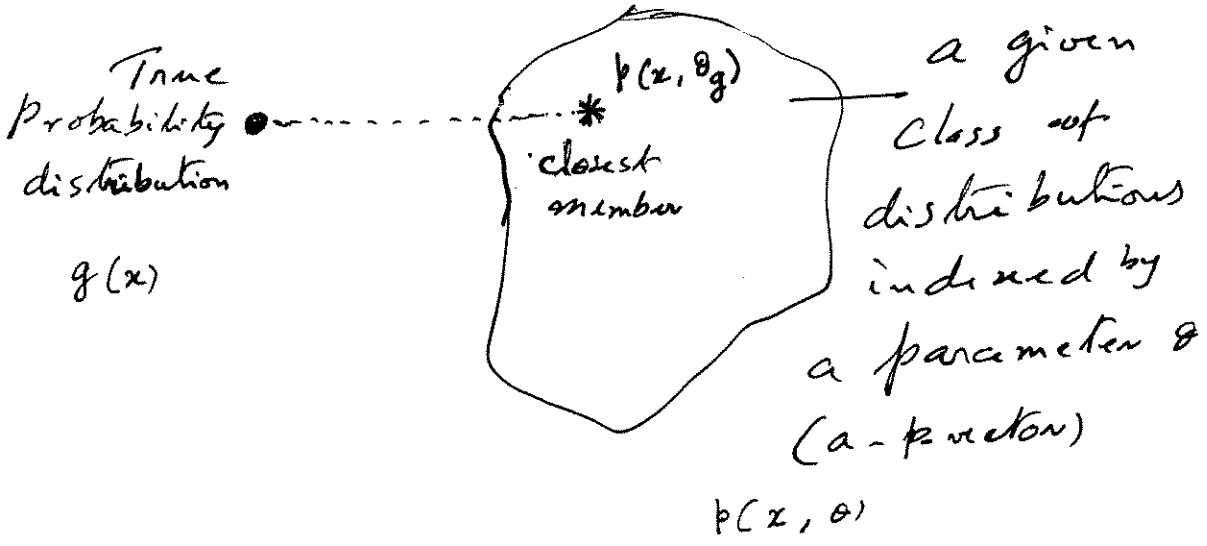
Proposed Methods:

- Cross validation using a loss function for prediction
- Information theoretic criteria minimizing a suitably chosen distance measure of the *true distribution* from a member in the *given set* of models

MODEL SELECTION

Information Theoretic Criteria

In practice, there is no particular method of choosing the class of probability distributions, a member of which is the parent of observed data. However, we can try to estimate the member in a given class of probability distributions which is the closest to the true distribution which may be outside the class.



θ_g is usually determined by minimizing some measure of distance between the probability densities g and $p(x,\theta)$. An appropriate measure is Kullback-Leibler measure of separation

$$\int g(x) \log[g(x)/p(x,\theta)]\, dx$$

$$= \int [g(x) \log g(x) - g(x) \log p(x,\theta)]\, d\theta$$

i.e. θ_g is then

$$\boxed{\theta_g = \arg\max_{\theta} \int g(x) \log p(x,\theta)\, dx}$$

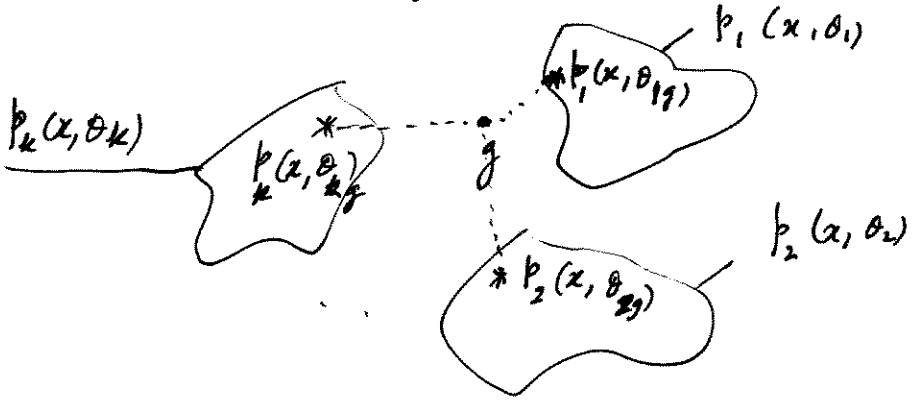
Our problem is as follows:

There are alternative classes

$$p_1(x,\theta_1),\ p_2(x,\theta_2),\ \ldots,\ p_k(x,\theta_k)$$

with corresponding members close to

$$p_1(x,\theta_{1g}),\ p_2(x,\theta_{2g}),\ \ldots,\ p_k(x,\theta_{kg})$$



The best approximation to g by a member of

$$\{ p_i(x, \theta_{ig}), i = 1, \ldots, k \}$$

is the one which maximises

$$KL(g, i) = \int g(x) \log p_i(x, \theta_{ig}) \, dx$$

with respect to i.

But we do not know g. An estimate of $KL(g, i)$ is

$$\widetilde{KL}(g, i) = \sum_{j=1}^{n} \log p_i(x_j, \hat{\theta}_{ig})$$

where $(x_1, \ldots, x_n)$ is the sample and

$$\hat{\theta}_{ig} = \arg \max_{\theta_i} \sum_{j=1}^{n} \log p_i(x_j, \theta_i)$$

Unfortunately $\widetilde{KL}(g, i)$ is not an unbiased estimate of $KL(g, i)$. Attempts have been made to make a correction to $\widetilde{KL}(g, i)$ so that $\widetilde{KL}(g, i)$ can be used in the place of $KL(g, i)$ in the Akaike decision rule ( )

1. Akaike Information Criterion

$$AIC(\theta_i) = 2\widetilde{KL}(g, i) - 2p$$

where p is the number of parameters estimated.

The rule overfits and is not consistent, but works well in practice. [Note that the usual convention is to take the negative of $AIC(\theta_i)$ and minimise w.r.t i]

2. Schwarz Information Criterion (or Bayes)

$$SIC\ (BIC) = 2\widetilde{KL}(g, i) - p \log n$$

The rule underfits but is strongly consistent

3. Minimum Description Length Criterion

$$MDL = \widetilde{KL}(g, i) + \frac{p}{2} \log\left(\frac{n}{2\pi}\right) + \log \int \sqrt{I(\theta_i)}\, d\theta_i$$

where $I(\theta_i)$ is the information matrix on θ_i based on the model $f_i(x, \theta_i)$.

The rule is strongly consistent.

General Information Criterion (strongly consistent)

$$GIC = \tilde{KL}(g,i) \Phi - \phi C_n$$

where

$$\lim_{n\to\infty} \frac{1}{n} C_n = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{C_n}{\log\log n} = +\infty$$

In each case the choice of the class $\Phi_i(x, \theta_i)$ is made by minimizing the criterion.

Some notes

These criteria are valid under heavy regularity conditions and the different alternative classes are nested in some sense.

The criteria do not take into account the purpose for which the model is estimated and the loss incurred in using the estimated distribution to predict future events.

One of the examples in which the astronomers are interested is to fit an appropriate mixture model

$$p_i(x, \underset{\sim}{\theta}) = \alpha_1 N(x : \mu_1, \Sigma_1) + \cdots + \alpha_i N(x ; \mu_i, \Sigma_i)$$

with $i = 1, \ldots, k$

where $N(x; \mu, \Sigma)$ is a multivariate normal density with mean μ and covariance matrix Σ. The alternative classes are the number of components in the mixture with k as the upper limit.

The criteria as developed are not strictly applicable as the mixture models do not satisfy the regularity conditions needed for their use. But simulation studies show that they are not very much affected by the lack of regularity conditions. However some caution is needed in using these criteria. Criteria based on other distance measures in the place of Kullback-Leibler are being developed for application in the

TABLE 1
Measurements at different time points

Individuals	Time Points				
	t_1	⋯	t_p	t_{p+1}	t_{p+2}
Past					
1	y_{11}	⋯	y_{p1}	$y_{p+1,1}$	$y_{p+2,1}$
⋮	⋮		⋮	⋮	⋮
n	y_{1n}	⋯	y_{pn}	$y_{p+1,n}$	$y_{p+2,n}$
Current					
c	y_{1c}	⋯	y_{pc}	?	?

The values to be predicted are indicated by ?

TABLE 2
Weights of 13 male mice measured at successive intervals of 3 days over 21 days from birth to weaning (Williams and Izenman, 1981)

	Day 3	Day 6	Day 9	Day 12	Day 15	Day 18	Day 21
1	0.190	0.388	0.621	0.823	1.078	1.132	1.191
2	0.218	0.393	0.568	0.729	0.839	0.852	1.004
3	0.211	0.394	0.549	0.700	0.783	0.870	0.925
4	0.209	0.419	0.645	0.850	1.001	1.026	1.069
5	0.193	0.362	0.520	0.530	0.641	0.640[a]	0.751
6	0.201	0.361	0.502	0.530	0.657	0.762	0.888
7	0.202	0.370	0.498	0.650	0.795	0.858	0.910
8	0.190	0.350	0.510	0.666	0.819	0.879	0.929
9	0.219	0.399	0.578	0.699	0.709	0.822	0.953
10	0.225	0.400	0.545	0.690	0.796	0.825	0.836
11	0.224	0.381	0.577	0.756	0.869	0.929	0.999
12	0.187	0.329	0.441	0.525	0.589	0.621	0.796
13	0.278	0.471	0.606	0.770	0.888	1.001	1.105

[a] This could be a recording error, but no change was made in the present computations.

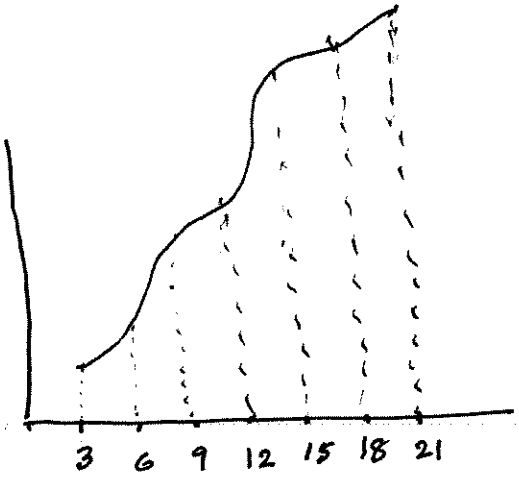
TABLE 3
Ramus heights of 20 boys at different ages (Elston and Grizzle, 1962; Grizzle and Allen, 1969; Lee and Geisser, 1975)

	8 yr	8½ yr	9 yr	9½ yr
1	47.8	48.8	49.0	49.7
2	46.4	47.3	47.7	48.4
3	46.3	46.8	47.8	48.5
4	45.1	45.3	46.1	47.2
5	47.6	48.5	48.9	49.3
6	52.5	53.2	53.3	53.7
7	51.2	53.0	54.3	54.5
8	49.8	50.0	50.3	52.7
9	48.1	50.8	52.3	54.4
10	45.0	47.0	47.3	48.3
11	51.2	51.4	51.6	51.9
12	48.5	49.2	53.0	55.5
13	52.1	52.8	53.7	55.0
14	48.2	48.9	49.3	49.8
15	49.6	50.4	51.2	51.8
16	50.7	51.7	52.7	53.3
17	47.2	47.7	48.4	49.5
18	53.3	54.6	55.1	55.3
19	46.2	47.5	48.1	48.4
20	46.3	47.6	51.3	51.8

TABLE 4
Dental measurements of 11 girls and 16 boys (Potthoff and Roy, 1964; Lee and Geisser, 1975)

	8 yr	10 yr	12 yr	14 yr
Girls				
1	21	20	21.5	23
2	21	21.5	24	25.5
3	20.5	24	24.5	26
4	23.5	24.5	25	26.5
5	21.5	23	22.5	23.5
6	20	21	21	22.5
7	21.5	22.5	23	25
8	23	23	23.5	24
9	20	21	22	21.5
10	16.5	19	19	19.5
11	24.5	25	28	28
Boys				
12	26	25	29	31
13	21.5	22.5	23	26.5
14	23	22.5	24	27.5
15	25.5	27.5	26.5	27
16	20	23.5	22.5	26
17	24.5	25.5	27	28.5
18	22	22	24.5	26.5
19	24	21.5	24.5	25.5
20	23	20.5	31	26
21	27.5	28	31	31.5
22	23	23	23.5	25
23	21.5	23.5	24	28
24	17	24.5	26	29.5
25	22.5	25.5	25.5	26
26	23	24.5	26	30
27	22	21.5	23.5	25

Prediction of mouse weight on day 21



$$y_{21} = a_1 + b_{3,1} y_3 + \cdots + b_{18,1} y_{18} + \epsilon_1$$

$$y_{21} = a_2 + b_{6,2} y_6 + \cdots + b_{18,2} y_{18} + \epsilon_2$$

$$\vdots$$

$$* \quad y_{21} = a_6 + b_{18,6} y_6 + \epsilon_6$$

are possible models.

LOO (Leave one out method)

Estimate the regression coefficients on 12 mice and use them to estimate y_{21} based on the earlier measurements on the 13th mouse.

The best model is the prediction of y_{21} based on the previous measurement y_{18} only. Does not depend on the growth process. However if we want the prediction of the y value at say 10 days, we may have to use a different model

TABLE 7

CVAE of different predictors under the polynomial growth curve model

Previous measurements used (1)	Degree of polynomial fitted (2)	Individual regression predictor (3)	Regression on polynomial coefficients (4)	Calibrated predictor of column 3 (5)	Empirical Bayes predictor (6)
Mice data (prediction of Y_7, $n = 13$)[a]					
Y_1–Y_6	5	7.472	.095	.252	
	4	.600	.076	.235	.375
	3	.175	.058	.093	.139
	2	.104	.060	.037	.087
	1	.206	.049	.035	.194
Y_2–Y_6	4	2.405	.079	.235	
	3	.241	.064	.141	.174
	2	.095	.040	.040	.075
	1	.158	.043	.035	.143
Y_3–Y_6	3	.757	.047	.192	
	2	.096	.039	.052	.069
	1	.111	.039	.034	.097
Y_4–Y_6	2	.229	.037	.094	
	1	.066	.036	.034	.054
Y_5–Y_6	1	.055	.031	.033	
Ramus data (prediction of Y_4, $n = 20$)					
Y_1–Y_3	2	2.989	.769	2.172	
	1	.584	.716	.638	.498
Y_2–Y_3	1	.812	.577	.751	
Dental data (prediction of Y_4, $n = 27$)					
Y_1–Y_3	2	47.398	4.430	9.483	
	1	3.998	3.288	3.680	2.322
Y_2–Y_3	1	12.426	3.585	8.358	

[a] Entries are 13 times the actual values.

Cross-validation assessment error of simple linear regression predictor

Previous measurements used	Direct regression	Inverse regression
Mice data (prediction of Y_7, $n = 13$)[a]		
Y_1–Y_6	.095	.103
Y_2–Y_6	.079	.081
Y_3–Y_6	.047	.048
Y_4–Y_6	.037	.040
Y_5–Y_6	.031	.034
Y_6	.027	.028
Ramus data (prediction of Y_4, $n = 20$)		
Y_1–Y_3	.769	.808
Y_2–Y_3	.577	.608
Y_3	.566	.618
Dental data (prediction of Y_4, $n = 27$)		
Y_1–Y_3	4.430	6.211
Y_2–Y_3	3.588	5.227
Y_3	3.665	4.929

[a] The entries are 13 times actual values.

REFERENCES

Tests of normality

Methods for Statistical Data Analysis of Multivariate Observations
R. Gnanadesikan, John Wiley
Pages 189-226

Tests on a circle and Sphere (randomness)

Circular Statistics in Biology
Edward Batschelet, Academic Press

Chi-Square, Likelihood ratio, Wald and score tests (Holy Trinity)

Linear Statistical Inference and its Applications
C. R. Rao, John Wiley
Chapters 6

Kolmogorov, Cramer-von Mises Test of goodness-of-fit

Elements of Large Sample Theory
E. Lehman, Springer
Pages 343, 397, 342, 344

Model Selection

Model Selection and Inference
A practical information theoretic approach
by Kenneth P. Burnham and David R. Anderson,
Springer, 1998

Model Selection, IMS Lecture Notes Volume 38
2001