Approximating data and statistical procedures. I. Approximating data *

P. L. Davies

March 25, 2003

Abstract

Stochastic models approximate data and are not true representations of the same. Statistical procedures make use of approximate stochastic models to facilitate the analysis of data.

1 Rationality

D. W. Müller (1974)

> ... die distanzierte Rationalität. Damit ist ein Verhalten gegenüber Sachgegebenheiten gemeint, das sich nicht von deren etwaigen oder vermeintlichen Eigengesetzlichkeiten leiten läßt, sondern ihnen mit Entwürfen des Verstandes in der Form von Modellen, Hypothesen, Arbeitshypothesen, Definitionen, Folgerungen, Alternativen, Analogien, also sozusagen "aus der Distanz", in der Weise partiellen, vorläufigen, approximativen Wissens, gegenübertritt.

2 The lack of a concept of approximation

Although statistical models are largely ad hoc with little if any theoretical underpinning statistical theory is based on the premise of true models. This will be denied by many but can be confirmed by taking any book on statistics and looking up the word "approximation" in the index. Statistics has no concept of an approximate stochastic model.

*Research supported in part by Sonderforschungsbereich 475, University of Dortmund

3 Approximation

A stochastic model P is an adequate approximation for a given data set $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ if a "typical" sample $\boldsymbol{X}_n(P) = (X_1(P), \ldots, X_n(P))$ generated under the model P "looks like" the data $\boldsymbol{x}_n$.

4 Looks like

The words "looks like" are operationalized by specifying certain features of the sample which are of interest. They may be chosen on the basis of substantive knowledge of the data or on properties of the simulated data sets $\boldsymbol{X}_n(P)$. Certain features of simulated data sets may not be granted legitimacy when comparing simulated and real data sets. Tukey (1993c)

> ...we should have to say that certain aspects of the data – not typically, but unavoidably, including "Most (Modelled) observations have irrational values !" – are not to be used in relating conceptual (or simulated) samples to observed samples. Thought and debate as to just which aspects are to be denied legitimacy will be both necessary and valuable.

5 Features not feature

The use of the plural in Section 4 was intentional. In general a model will be adequate only if the simulated samples exhibit several different features which are judged to be relevant. There is no general principle of defining what is adequate, it depends on the circumstances.

6 Typical

The word "typical" is quantified by specifying a number $\alpha, 0 < \alpha < 1$, such that at least $100\alpha\%$ of simulated samples exhibit the features of interest.

7 An illustrative experiment

Using P generate 999 samples $\boldsymbol{X}_{i,n}(P)$, $i = 1, \ldots, 999$ each of size n and then insert the real data set $\boldsymbol{x}_n$ at random. Specify α and then name $1000(1-\alpha)$ "untypical" data sets. If the real data set is one of those named then the model is not an adequate approximation. Small versions are given by Figures 1, 2 and 3. The real data sets of Figures 1 and 3 were kindly provided by Prof. Dieter Mergel of the Physics Department of the University of Essen. They relate to the intensity of reflected X-rays as a function of the angle of incidence and come from the area of thin film physics. The real data set of Figure 2 gives the daily rates of return for a financial index.

8 Algorithms

The operationalization of "looks like" and the quantification of "typical" will in principle result in an algorithm with inputs P and $\boldsymbol{x}_n$ which will determine whether or not the model P is an adequate approximation for the data $\boldsymbol{x}_n$.

9 Direct comparison

The concept of approximation involves only the postulated model P via its samples $\boldsymbol{X}_n(P)$ and the data $\boldsymbol{x}_n$. It is a direct comparison of samples generated under P and $\boldsymbol{x}_n$. It does not involve a comparison of P with some postulated "true" generating mechanism Q of $\boldsymbol{x}_n$.

10 Approximation for the data at hand

A model P is an adequate approximation or not for the data at hand $\boldsymbol{x}_n$. There is no assumption that $\boldsymbol{x}_n$ is embedded in a sequence of data sets as is done in the frequentist interpretation. The degree of belief in the truth of any model is zero. Treating stochastic models as approximations to given data sets is neither frequentist nor subjective.

11 Data generated by the model

If the data $\boldsymbol{x}_n$ were in fact generated by the model P (there is no assumption that this is the case) then the model P will be adequate for $\boldsymbol{x}_n$ with probability at least α.

12 Adequacy regions

Given a data set $\boldsymbol{x}_n$ and a family of models $\{P_\theta : \theta \in \Theta\}$ we define the approximation region for $\boldsymbol{x}_n$ by

$$\mathcal{A}(\boldsymbol{x}_n, \Theta) = \{\theta : P_\theta \text{ is an adequate approximation for } \boldsymbol{x}_n\} \tag{1}$$

The approximation region $\mathcal{A}(\boldsymbol{x}_n, \Theta)$ may be empty. If $\boldsymbol{x}_n$ is equal to $\boldsymbol{X}_n(\theta_0)$ for some $\theta_0 \in \Theta$ then $\theta_0 \in \mathcal{A}(\boldsymbol{x}_n, \Theta)$ with probability at least α. If it is known that $\boldsymbol{x}_n$ was so generated then $\mathcal{A}(\boldsymbol{x}_n, \Theta)$ is an α-confidence region for θ_0. This will only be the case in simulations where, ignoring problems of randomness for the moment, data can be generated according to a model. In the frequentist approach confidence intervals are often justified by claiming that they will contain the true parameter in say 95% of the cases for a sequence of real data sets (see for example Bickel and Doksum (1977), page 163). In the adequacy interpretation for any $\theta \in \mathcal{A}(\boldsymbol{x}_n, \Theta)$ 95% of the data sets generated und P_θ will look like $\boldsymbol{x}_n$.
If $\mathcal{A}(\boldsymbol{x}_n, \Theta)$ is an adequacy region and $T(\theta)$ is some functional then the adequacy region for T is $\{T(\theta) : \theta \in \mathcal{A}(\boldsymbol{x}_n, \Theta)\}$.

13 An example

We consider a sample $\boldsymbol{x}_n$ of size n and the family of models is the normal family $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$. We denote the standard deviation of the sample by s_n and the mean by $\bar{\boldsymbol{x}}_n$. The model $\mathcal{N}(\mu, \sigma^2)$ will be considered as an adequate approximation for the data if the following hold:

$$\mathrm{qchi2}(0.01, n-1) \leq \; s_n^2/\sigma^2 \; \leq \mathrm{qchi2}(0.99, n-1) \tag{2}$$

$$|(\bar{\boldsymbol{x}}_n - \mu)/\sigma| \leq z(0.99)/\sqrt{n} \tag{3}$$

$$d_{ko}(F_n, \mathcal{N}(\mu, \sigma^2)) \leq \mathrm{qdko}(0.01, n) \tag{4}$$

where qchi2(α, k) denotes the $\alpha-$quantile of the chi-squared distribution with k degrees of freedom, $z(\alpha)$ denotes the α-quantile of the standard normal, d_{ko} denotes the Kolmogoroff metric, F_n is the empirical distribution function associated with $\boldsymbol{x}_n$ and qdko(α, n) is the $\alpha-$quantile of the Kolmogoroff metric for a sample of size n.
The corresponding approximations for a normal sample (right panel) and a double exponential sample (left panel) both of size 50 are shown in Figure 4. The samples were standardized to have mean 0 and standard deviation 1.

14 Precision

In the case of continuous probability models the property that simulated samples are irrational with probability one will be denied legitimacy (Section 4). The simulated samples will be truncated to the precision of the data $\boldsymbol{x}_n$.

15 Generating i.i.d. univariate samples

Suppose we have a real sample $\boldsymbol{x}_n$ and a model involving independently and identically distributed random variables with a continuous distribution function F. In accordance with Section 14 the samples of size n generated using F are truncated to the precision of $\boldsymbol{x}_n$, say ε. Given any $\delta > 0$ we can consider a distribution G such that the Kolmogoroff distance between F and G,

$$d_{ko}(F, G) = \sup_x |F(x) - G(x)| \tag{5}$$

satisfies $d_{ko}(F, G) < \delta$. If δ is sufficiently small and we generate the samples by

$$X_i(F) = F^{-1}(U_i), \quad X_i(G) = G^{-1}(U_i),\ i = 1, \ldots, n \tag{6}$$

where the U_i are i.i.d uniform on [0, 1] then the truncated samples will be equal with high probability. Consequently if F is an adequate model for $\boldsymbol{x}_n$ then so is G.

16 Topologies, weak and strong

The topology of data analysis and hence of all of statistics is a weak topology. We use the term "weak" in its customary topological sense. On identifying a topology with its open sets a topology $\mathcal{O}_1$ is weaker that a topology $\mathcal{O}_2$ if $\mathcal{O}_1 \subset \mathcal{O}_2$ (Hewitt and Ross (1979) page 9). A weak topology is not to be confused with *the* weak topology although *the* weak topology is *a* weak topology (see Donoho and Liu (1988)). Typical weak topologies are generated by metrics of the form

$$d_{\mathcal{C}}(P,Q) = \sup\{|P(C) - Q(C)| : C \in \mathcal{C}\} \tag{7}$$

where $\mathcal{C}$ is a Vapnik-Cervonenkis class of (Borel) subsets of $\mathbb{R}^n$ (see for example Pollard (1984), Vapnik (1998)). The Kolmogoroff metric (5) is a weak metric. Strong metrics are density based and are related to the total variation metric

$$d_{tv}(P,Q) = \sup\{|P(B) - Q(B)| : B \text{ Borel}\}. \tag{8}$$

The Hellinger metric is a strong metric, Kullback-Leibler is a strong discrepancy.

17 Topologies and approximation

Weak topologies are consistent with approximation whereas strong ones are not. Given a data set $\boldsymbol{x}_n$ and a continuous probability model P which is an adequate approximation for the data there will be other probability models Q which are also adequate approximations but with $d_{tv}(P,Q) = 1$. In the case of Section 15 this follows on noting that given F and given δ however small there are distributions G with $d_{ko}(F,G) < \delta$ but with $d_{tv}(F,G) = 1$. For large n it is perfectly reasonable to approximate the binomial distribution $bin(n, 0.5)$ by the normal distribution. This is not possible in a strong metric. Sometimes the discrete nature of quantity is not discernable at the macro level, atoms or the energy levels of a photon. A natural approximation by a continuous distribution is not possible in a strong metric.

18 Topologies and direct comparison

Given a continuous model P and a simulated data set $\boldsymbol{X}_n(P)$ we form the empirical measure $P_n(P) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i(P)}$ where δ_x denotes the unit measure in x. We have $d_{tv}(P_n(P), P) = 1$ whereas for a weak metric $d_{\mathcal{C}}$ based on a Vapnik-Cervonenkis class $d_{\mathcal{C}}(P_n(P), P) = \mathrm{O}(1/\sqrt{n})$ whatever P. Weak metrics permit direct comparison of the data and the model whereas strong metrics allow only an indirect comparison.

19 Topologies and the differential operator

Given an absolutely continuous distribution function F

$$F(x) = \int_{-\infty}^{x} f(u)\, du$$

the differential operator D is defined by

$$D(F) = f. \tag{9}$$

If a weak metric is used on the space of distribution functions $\mathcal{F}$ and some usual norm on the space of density functions $\mathcal{D}$ then the differential operator $D : \mathcal{F} \rightarrow \mathcal{D}$ is discontinuous at every point $F \in \mathcal{F}$.

20 Smooth functionals

A functional T defined on the space of probability distributions is smooth in different degrees if it is bounded or continuous or differentiable with respect to a weak metric. The gold standard is locally uniform Fréchet differentiability (Davies (1998), Bednarski and Clarke (1998)). As weak topologies have few open sets it is more difficult to construct such functionals than it is to construct functionals which are smooth with respect to strong topologies.

21 Asymptotics and direct comparison

The direct comparison of the real data $\boldsymbol{x}_n$ with simulated data sets $\boldsymbol{X}_n(P)$ can sometimes be simplified by the use of asymptotics. To be of use these must be locally uniform and supplementable by simulations for small sample sizes. Locally uniformly differentiable functionals in a weak metric (Section 20) give rise to locally uniform asymptotics.

22 Locally uniform Fréchet differentiability: an example

Consider an M-functional $T(P) = (T_L(P), T_S(P))$ defined by

$$\int \psi\left(\frac{T_L(P)-x}{T_S(P)}\right) dP(x) \;=\; 0 \tag{10}$$

$$\int \chi\left(\frac{T_L(P)-x}{T_S(P)}\right) dP(x) \;=\; 0. \tag{11}$$

Given a data set $\boldsymbol{x}_n$ we consider a probability model P as adequate approximation for $\boldsymbol{x}_n$ if it satisfies

$$d_{ko}(P_n, P) \le \mathrm{qdko}(\alpha_1, n) \tag{12}$$

and

$$\mathrm{q}((1-\alpha_2)/2, T, P, n) \le \frac{T_L(P_n) - T_L(P)}{T_S(P_n)} \le \mathrm{q}((1+\alpha_2)/2, T, P, n) \tag{13}$$

where P_n denotes the empirical probability based on $\boldsymbol{x}_n$ and $q(\alpha, T, P, n)$ denotes the q-quantile of $(T_L(P_n(P)) - T_L(P))/T_S(P_n(P))$. Davies (1998) shows that under suitable smoothness conditions on ψ and χ and under *weak* conditions on P which involve only its largest atom, T is locally uniformly Fréchet differentiable at P in the Kolmogoroff metric. In particular

$$|q(\alpha, T, P, n) - q(\alpha, T, Q, n)| \le C(P) d_{ko}(Q, P) \tag{14}$$

and we also have locally uniform convergence to a normal distribution with a quantifiable error term. Putting all this together we can approximate $q(\alpha, T, P, n)$ by $q(\alpha, T, P_n, n)$ for P satisfying (12) and $q(\alpha, T, P_n, n)$ itself by

$$z(\alpha)\Sigma(T, n, P_n)/\sqrt{n}$$

where

$$\Sigma(T, n, P)^2 = \frac{\int \psi\left(\frac{T_L(P)-x}{T_S(P)}\right)^2 dP(x)}{\left(\int \psi^{(1)}\left(\frac{T_L(P)-x}{T_S(P)}\right) dP(x)\right)^2} \tag{15}$$

and $z(\alpha)$ is the $\alpha-$ quantile of the standard normal distribution. The errors involved in these approximation are in principle calculable given the

data	location	scale	emp. int.	Gau. int
$\boldsymbol{xg}_{50}$	0.105	0.813	[-0.144,0.354]	[-0.145,0.355]
$\boldsymbol{xc}_{50}$	0.141	0.1.70	[-0.405,0.687]	[-0.380,0.622]

Table 1:

data $\boldsymbol{x}_n$. Without the local uniformity one could only appeal to non-uniform asymptotics.
We set

$$\psi(x) = \frac{\exp(x/5) - 1}{\exp(x/5) + 1} \tag{16}$$

$$\chi(x) = \frac{x^4 - 1}{x^4 + 1} \tag{17}$$

and apply the T-functional to two data sets of size $n = 50$. The first $\boldsymbol{xg}_{50}$ is i.i.d. standard Gaussian and the second $\boldsymbol{xc}_{50}$ is i.i.d. standard Cauchy. We set $\alpha_1 = 0.99$ and $\alpha_2 = 0.96$ to give $\alpha \geq 0.95$. The results are given in Table 1 where emp. int. and gau. int. refer respectively to the approximation intervals for $T_L(P)$ based on the quantiles derived from (15) with $P = P_n$ and $P = N(0, 1)$ respectively.

23 Existence and uniqueness

Often it is necessary to impose conditions on a probability measure P to ensure the existence and uniqueness of some functional T at P. For example if T is the functional of the Section 22 then Huber (1981) pages 138-139 requires only that the largest atom $\Delta(P)$ of P should satisfy $\Delta(P) < \chi(\pm\infty)/(\chi(\pm\infty) - \chi(0))$. In contrast Davies (1987) when considering the existence and uniqueness of so called S-functionals requires that P have a symmetric decreasing density f which is strictly decreasing at some point which depends on the function defining the S-functional. Such a condition is *strong*. There is an unfortunate tendency to dismiss a condition which requires, for example, only the existence of a continuous density as weak. One reads such expressions as "under weak conditions" or "under general conditions". This is to be deprecated. Such conditions are strong, not weak, and are not acceptable.

24 Residual based approximation and scale

Definitions of approximation based on residuals are often useful, particularly in regression problems. A source of difficulty is in determining the appropriate order of magnitude for the noise or the residuals. There is a tendency for L_2-methods to produce residuals which "look like" those required by the model even in situations where the model is not an adequate approximation. An example are outliers in the standard linear regression problem. In some nonparametric regression problems there is also a problem in determining an appropriate scale. Thought is necessary when addressing this problem.

25 Density based statistics

Many concepts in statistics such as likelihood, Hellinger distance, AIC and Minimum Description Length are density based. Because of Sections 17 and 19 an adequate theory of approximation cannot be derived from density based concepts. They have only a minor role to play in statistical theory and practice. They can probably be dispensed with.

26 Pathologies of maximum likelihood

Given an ordinary non-pathological univariate dataset $\boldsymbol{x}_n$, to be explicit, say a sample of size n from the standard Gaussian distribution $N(0,1)$ and given any M and any $\varepsilon > 0$ there exists a distribution F with the following properties:

(i) $d_{tv}(F, N(0,1)) < \varepsilon$

(ii) F is symmetric with mean zero and variance 1

(iii) F has a positive infinitely differentiable density f

(iv) F has a moment generating function, in particular moments of all orders are finite

(v) the maximum likelihood of μ based on the location family $F(\cdot - \mu)$ and the data $\boldsymbol{x}_n$ is M.

Because of (i) F is indistinguishable from the $N(0,1)$ distribution, (iii) and (iv) are irrelevant but look good. The maximum likelihood estimator based on the location model $F(\cdot - \mu)$ can be made to take on any desired value.

27 Pathologies of Bayes

Consider the sample of Section 26 with sample size $n \geq 2$. Let Γ be a prior distribution for μ with a positive, continuous and bounded density γ. Then given M, $\varepsilon > 0$, $\delta > 0$ and $\eta > 0$ there exists a distribution F satisfying (i)-(iv) of Section 26 such that the posterior distribution $\Delta(\cdot|\boldsymbol{x}_n)$ satisfies

$$\Delta((M-\delta, M+\delta)|\boldsymbol{x}_n) \geq 1-\eta. \tag{18}$$

This corresponds to the result of Section 26. The posterior distribution for μ can be made to be arbitrarily highly concentrated on an arbitrarily small interval about an arbitrary point.

28 Pathologies of efficiency

Consider the sample of Section 26 with sample size $n \geq 2$. Then given $\varepsilon > 0$ and $\delta > 0$ there exists a distribution F satisfying (i)-(iv) of Section 26 such that the 95%-confidence interval $I(\boldsymbol{x}_n)$ for μ based on the location family $F(\cdot - \mu)$ and the data $\boldsymbol{x}_n$ satisfies $I(\boldsymbol{x}_n) \subset (-\delta, \delta)$. In other words the confidence interval can be made arbitrarily small using a distribution which is empirically indistinguishable from the normal distribution with the same mean and variance.

29 Data independent pathologies

The pathologies of Sections 26 and 27 rely on constructing a data dependent model. Given the finiteness of precision of all real data similar constructions will give rise to data independent pathologies of a similar form.

30 Commonsense and pathologies

The location models of Sections 26, 27 and 28 could be dismissed on commonsense grounds as being silly, absurd or pathological. Firstly, this is not the case. The models are not pathological as data generated by them are indistinguishable from normal data for any practical degree of precision. Secondly, such adjectives are of no help in understanding the problem. The pathologies arise only in conjunction with the procedures "prescribed" by the model. It is the combination of procedures and model which gives rise to the pathologies. There is no reason for allowing models to prescribe anything.

31 Likelihood, sufficiency and perturbations

Given a parametric model $P_\theta, \theta \in \Theta$ and a sample $\boldsymbol{x}_n$ the likelihood principle states that the likelihood function contains all relevant information about θ contained in the data (Berger (1980) pages 23-28). For certain parametric likelihood functions a factorization is possible which defines a sufficient statistic. In this case the sufficient statistic contains all relevant information about θ contained in the sample. For i.i.d. Gaussian sample $\boldsymbol{X}_n(\theta)$ the mean and standard deviation form a sufficient statistic. They do not contain all the relevant information about θ contained in the data. They do not contain the most important information, namely that the data may be well approximated by a Gaussian model with appropriate parameter values. The small print contains the proviso that the likelihood principle only holds for data which are distributed according to P_θ for some θ but even this is not sufficient. Not only must the data be so distributed but we must also know that they are so distributed. This brings us back to Section 2. The likelihood principle is based on revealed truth.

Likelihood requires a density with respect to a single measure which dominates all the distributions $P_\theta, \theta \in \Theta$. In continuous models there exist arbitrarily small perturbations $Q_\theta, \theta \in \Theta$ which are not dominated by a single measure and hence for which no likelihood is available. Even if we restrict considerations to perturbations with likelihood these can, in conjunction with the likelihood principle, be chosen to give almost any pathological result desired. Similar considerations hold for the concept of sufficiency. Likelihood and sufficiency are both pathologically discontinuous in the weak topology of data analysis. Given data and a model with a likelihood function there is no reason for basing the analysis on this function and every reason for not doing so.

32 Blandness: pathologies tamed

Section 28 shows the possibility of importing precision via the model by means of model optimal procedures. This can be avoided by using smooth functionals (Section 20) or to a limited extent by using models which are "bland" or "hornless" (Tukey (1993c)). Any peculiarity or "horn" of a model will be exploited to the fullest extent by an optimal procedure. The Cauchy distribution has a slight horn, namely its peakedness near the origin. Bland models are useful for comparison. We are fortunate that the normal distribution is bland. Tukey (1993c):

> NO ONE HAS EVER SHOWN THAT HE OR SHE HAD A FREE LUNCH
> Here, of course, "FREE LUNCH" means "usefulness of a model that is locally easy to make inferences from".

33 Why the normal distribution?

The normal distribution and optimal procedures based on it are often perfectly reasonable. The reason is that it is very difficult to estimate the population mean on the basis of a normal sample.

34 Regularization and pathologies

Statistical problems which involve continuous probability models are badly posed. The problem is not removed by requiring the model to be a reasonable approximation to the data by means, for example, of a goodness-of-fit test. General principles such as Bayes or likelihood can always be made to produce pathologies for models which are consistent with the data. These cannot be removed by qualitative assumptions such as smooth densities and finite moments (Sections 26, 27 and 28). To prevent pathologies the problem must be regularized or almost regularized. The normal distribution may be regarded as a regularization of the location problem where the regularization is done in terms of Fisher information (Section 33). Other forms of regularization are possible and are of interest (Huber (1981)). However it is as well to be aware of the inexplicit requirement of regularization behind some so called general principles of statistics.

35 Reasonable and pathological

There is a continuum between what is reasonable and what is pathological. It is not always easy to decide where we are on this scale.

36 Robustness

Our attitude to robustness is based on that of Huber (1981) whose approach is a functional analytical one. Robustness is or should be concerned with boundedness, continuity and differentiability of functionals in weak topologies. Because of its emphasis on efficiency the approach to robustness based on influence functions (see for example Section 2. of Hampel, Rousseeuw, Ronchetti and Stahel (1985)) inherits the problems of likelihood. Care must be taken with the basic model to avoid pathologies as in Section 28. The authors are aware of this (page 413 of Hampel et al (1986)) and their concept of a simple model is related to the concept of blandness (Section 32).

37 Model choice

Many stochastic models are indexed by an infinite dimensional parameter or by a sequence of finite dimensional parameter spaces of increasing dimension. Given a data set $\boldsymbol{x}_n$ there may be many models which are an adequate approximation for the data or there may be none. Procedures governing the choice of models must be able to deal with both situations and will in general also incorporate substantive knowledge about the data. A theory of model choice which does not allow the conclusion that none of the offered models is adequate is itself not adequate.

38 Model choice and simplicity

If a family of models is indexed by an infinite dimensional parameter then there can be infinitely many adequate models for a given data set $\boldsymbol{x}_n$. In such cases a concept of simplicity is required. An example is the regression function f in the nonparametric regression model

$$Y(t) = f(t) + \varepsilon(t), \quad 0 \le t \le 1. \tag{19}$$

Consider a model based on some infinitely differentiable f which is an adequate approximation to the data. We may perturb f in some manner to give rise to a function f' which is not even continuous but such that $\|f - f'\|_\infty < \varepsilon$ where ε is so small as to be non-observable. In this and similar situations we can only try and specify a maximum degree of smoothness which is compatible with an adequate approximation (Donoho (1988)). More generally we can only specify a minimum degree of complexity required for an adequate approximation. The operational definition of complexity may depend both on substantive knowledge and the theoretical properties of the models.

39 Model choice and substantive knowledge

Another approach to the choice of model which is related to the consideration of the previous section is one that is at least partially based on prior quantitative or substantive knowledge. Examples are a first derivative of at most 1.5 or monotone increasing or at most one local maximum. Prior knowledge of the form, a finite second moment, a continuous second derivative are not quantitative and can not help in choosing a model. See the comment to (iii) and (iv) of Section 26. Quantitative bounds on the function class involved lead to quantitative entropy bounds which appear in the exponential inequalities of empirical process theory (see Vapnik (1998)). These inequalities can be used when choosing a model. Classes of functions defined only by qualitative bounds have infinite entropy and are of no help when choosing a model.

40 Model choice and universal principles

Some procedures of model choice are universal in that they are based on a principle which is independent of any substantive knowledge about the data. Examples are Akaike's AIC (Akaike (1973, 1974, 1977, 1978 1981)), Schwarz's BIC (Schwarz (1978)), Bozdogan's ICOMP (Bozdogan (200)), Rissanen's MDL (Rissanen (1987)) and Bayes. Substantive knowledge is incorporated in the choice of the family of models and, in the Bayesian scheme, in the prior distribution. This having been done no further use is made of it. The choice of the model within the family is based on the universal principle. Each model is assigned a single real number which purports to represent the degree of fit of the model to the data, possibly modified to take the complexity of the model into account. That model is chosen which gives the best fit or, in the case of Bayes, the numbers represent some measure of the degree of belief in the model.

41 Model choice and sufficiency

Rissanen (1987)

> It is argued that all the useful information in observed data that can be extracted with a selected class of modelled distributions, will be obtained if we calculate the stochastic complexity, defined to be the shortest description length of the data.
>
> $\cdots * * * \cdots$
>
> Hence, if we could determine a model with which the stochastic complexity is reached, we would have learned all the useful information in the data that on the whole can be extracted with the chosen class, and that the only way anyone could teach us more is to offer a better model class. For this reason such a model may be regarded to represent the *sufficient statistic* [my italics] for the data ....

42 Universal principles, approximation and new models

The universal principles of Section 40 measure the (modified) goodness-of-fit of a model by a single number. Typically a definition of approximation of a data set by a model will depend on several features of both the data and the model, at least partly determined by substantive knowledge. It will not be reduceable to a single number (Section 5). In particular it is not possible to decide on the basis of the single numbers and the parameter values yielded by universal principles whether or not the model is an adequate approximation to the data or not. Thus irrespective of how bad a family of models is, universal principles of model selection give us no cause to look for a better family. Robustifying a universal principle does not weaken the above criticism.

43 Model choice: Nonparametric regression

This section is based on Davies and Kovac (2001). We consider data of the form $(t_i, y(t_i)), i = 1, \ldots, n$ with the t_i ordered points in [0, 1] and the $y(t_i)$ real valued. We look for an adequate approximation in the class of models $\{P(f, \sigma) : f : [0, 1] \to \mathbb{R}, \sigma > 0\}$ where random variables generated under $P(f, \sigma)$ are of the form

$$Y(t) = f(t) + \sigma\varepsilon(t) \tag{20}$$

with $\varepsilon(t)$ representing standard Gaussian white noise. In many situations noise is indeed well approximated by Gaussian white noise but the method can be extended to accommodate for example Cauchy noise. We treat σ as a nuisance parameter and set

$$s_n = 1.48\text{median}(|y(t_2) - y(t_1)|, \ldots, |y(t_n) - y(t_{n-1})|)/\sqrt{2}. \tag{21}$$

Approximation intervals for σ can also be given if desired. For any given f and any interval $I \subset \{1, \ldots, n\}$ we set

$$w(I) = \sum_{j \in I} (y(t_j) - f(t_j))/\sqrt{|I|} \tag{22}$$

where $|I|$ denotes the number of elements of I. The model based on f is an adequate approximation to the data if

$$\sup_{I \in \mathcal{I}} |w(I)| \le s_n \sqrt{\tau(\mathcal{I}) \log(n)} \tag{23}$$

where $\mathcal{I}$ is a family of subintervals of $\{1, \ldots, n\}$. If $\mathcal{I}$ is a multiresolution scheme with factor 2 we use the default value $\tau = 2.5$. The equations (21), (22) and (23) define the concept of approximation which we use. As a measure of complexity we take the number of local extremes of the function f. The final approximation problem is then to determine a function f with the minimum number of local extremes subject to (23). In Davies and Kovac (2001) developed a technology for solving this problem which is based on the taut string due to Hartigan and Hartigan (1985). Recently there has been progress on solving the problem directly but here we still use the taut string. The solution for the real dataset of Figure 3 is shown in Figure 5. The remaining data sets of Figure 3 were generated using the function of Figure 5 contaminated with Gaussian white noise with the appropriate variance and the final result being rounded to an integer.

44 Model choice: densities

Given a data set $\boldsymbol{x}_n$ we look for a density f which is an adequate approximation for the data i.e. data generated under the density will look like the original data set. The following is based on Davies and Kovac (2003). The notion of complexity we use is that of Section 43. The notion of adequacy we use is based on Kuiper metrics which are defined as follows. The Kuiper metric d_{ku}^k of order k is given

$$d_{ku}^k(F,G) = \sup \left\{ \sum_{i=1}^{k} (F(b_i) - F(a_i) - G(b_i) + G(a_i)) : a_1 < b_1 < \ldots a_k < b_k \right\} \tag{24}$$

We take an Kolmogoroff ball of radius ε centred at the empirical distribution function F_n and calculate the taut string TS through it with $TS(x_{(1)}) = 0$ and $TS(x_{(n)}) = 1$ where $x_{(1)} < \ldots < x_{(n)}$ are the order statistics. The radius of the ball is decreased until $d_{ku}^k(F_n, TS) \leq qu(n,k)$. For $n \geq 50$ we use the default value $k = 19$ and $qu(n,k)$ is such that for a uniform sample the probability that the taut string has modality 1 is 1/2. We note again that no mention is made of a "true" density and that the concept of approximation is based on a weak metric.

45 Bayes and learning

Bayesian statistics is conservative in the extreme. It is not possible to learn anything within the Bayesian framework: to learn you have to abandon it.

46 Bayes, bets and truth

A Bayesian prior can be interpreted as a degree of believe in the truth of a model. Just as in a horse race the Dutch book argument or, in more modern terms, an arbitrage argument shows that the betting odds must be describable by an additive probability measure. The additivity of the Bayesian prior rests on the assumption that it is not possible for two different models to be simultaneously true. If we bet on the number of black balls in an urn then it is not possible for the urn to simultaneously contain exactly five and exactly six black balls. If one could ascertain the truth of a statistical model then I suppose that in one sense the Bayesian approach could be justified. Simultaneously however it would be the end of Bayesian statistics as we would probably find that all our models are false. Bayesians bets are non-callable and this is indispensable for Bayesian statistics which requires both the idea of truth and the impossibility of ascertaining it.

47 Callable bets

In spite of the arguments of Section 46 there is a simple way of betting on models and parameter values such that the bets are callable. We do not bet that the model P_θ is the true model for the data $\boldsymbol{x}_n$ but rather that the model P_θ is an adequate approximation for the data $\boldsymbol{x}_n$. The definition of an adequate approximation will be operationalized by an algorithm (Section 8) and so one has only to run the algorithm to see whether the model is indeed an adequate approximation. We can quote odds for every model or parameter value but it is clear that these odds cannot be described by a probability distribution over the parameter space. Indeed, if for example I quote odds 50% that the $N(0,1)$ distribution is an adequate model then I will also quote odds of 50% for each of the models $N(10^{-100},1)$, $N(0,1+10^{-100})$ and $N(10^{-100},1+10^{-100}),\ldots$.

48 Determinism and randomness

It is often thought that statisticians analyse random data, a claim supported by much of the statistical literature where data sets are referred to as realizations of random variables. In spite of this no data set has ever been shown to be random nor is there any known method for generating random data. On the other hand there are deterministic data sets such as the decimal expansion of π which show all the characteristics of random data sets (Murier and Rousson (1998)). Stochastic methods can be applied with success to deterministic sequences such as that the distribution of the prime numbers (Kac (1959)). Any satisfactory theory of approximation must not place any assumptions of the randomness or otherwise of the data sets being analysed. The above arguments could be countered by claiming that the most basic scientific theory we have at present namely quantum mechanics has an inescapable random component in the random collapse of the wave packet.

This is however not the case. A theory of quantum mechanics going back to David Bohm is a deterministic one (Bell (1987), Albert (1992), Goldstein (1998a), (1988b), Dürr (2001)).

49 Tukey and approximation

An early version of this paper was sent to John Tukey who replied with three unpublished (Tukey (1993a), (1993b), (1993c)) articles which are now available from my website.

> Davies's emphasis on approximation is well-chosen and surprisingly novel. While these [sic] will undoubtedly be a place for much careful work in learning how to describe the concept - - and its applications - - in detail, it is clear that Davies has taken the decisive step by asserting that there must be a formal admission that adequate approximation, of one set of observable (or simulated) values by another set, needs to be treated as a practical identity.

50 Principles

Statistics has no principles. The mind is free tempered only by the reality of the world we live in (see Section 1).

References

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *Second international symposium on information theory*, pages 267–281, Budapest.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

[Akaike, 1977] Akaike, H. (1977). On entropy maximization principle. In Krishnaiah, P., editor, *Applications of Statistics*, pages 27–41. North Holland, Amsterdam.

[Akaike, 1978] Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika*, 65:53–59.

[Akaike, 1981] Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16:3–14.

[Albert, 1992] Albert, D. Z. (1992). *Quantum Mechanics and Experience.* Harvard University, Cambridge, Massachusetts.

[Bednarski and Clarke, 1998] Bednarski, T. and Clarke, B. R. (1998). On locally uniform expansions of regular functionals. *Discussiones Mathematicae: Algebra and Stochastic Methods*, 18:155–165.

[Bell, 1987] Bell, J. S. (1987). *Speakable and unspeakable in quantum mechanics.* Cambridge University press, Cambridge.

[Berger, 1980] Berger, J. O. (1980). *Statistical Decision Theory: Foundations, Concepts and Methods.* Springer, New York, Berlin, Heidelberg.

[Bickel and Doksum, 1977] Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics.* Holden-Day.

[Bozdogan, 2000] Bozdogan, H. (2000). Akaike's information criterion and recent developments in informtaion complexity. *Journal of Mathemtical Psychology*, 44:62–91.

[Davies, 1987] Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics*, 15:1269–1292.

[Davies, 1998] Davies, P. L. (1998). On locally uniformly linearizable high breakdown location and scale functionals. *Annals of Statistics*, 26:1103–1125.

[Davies and Kovac, 2001] Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Annals of Statistics*, 29(1):1–65.

[Davies and Kovac, 2002] Davies, P. L. and Kovac, A. (2002). Densities, spectral densities and modality. Technical Report 53, SFB 475, University of Dortmund, Dortmund, Germany.

[Donoho and Liu, 1988] Donoho, D. and Liu, R. (1988). Pathologies of some minimum distance estimators. *Annals of Statistics*, 16(2):587–605.

[Donoho, 1988] Donoho, D. L. (1988). One-sided inference about functionals of a density. *Annals of Statistics*, 16:1390–1420.

[Dürr, 2001] Dürr, D. (2001). *Bohmsche Mechanik als Grundlage der Quantenmechanik*. Springer, Berlin, Heidelberg.

[Goldstein, 1998a] Goldstein, S. (1998a). Quantum theory without observers- part one. *Physics Today*, pages 42–46.

[Goldstein, 1998b] Goldstein, S. (1998b). Quantum theory without observers- part two. *Physics Today*, pages 38–42.

[Hampel et al., 1986] Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

[Hartigan and Hartigan, 1985] Hartigan, J. A. and Hartigan, P. M. (1985). The dip test of unimodality. *Annals of Statistics*, 13:70–84.

[Hewitt and Ross, 1979] Hewitt, E. and Ross, K. A. (1979). *Abstarct Harmonic Analysis I*. Springer, Berlin/Heidelberg.

[Huber, 1981] Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

[Kac, 1959] Kac, M. (1959). *Statistical Independence in Probability, Analysis and Number Theory*. Number 12 in The Carus Mathematical Monographs. Wiley.

[Müller, 1974] Müller, D. W. (1974). Thesen zur Didaktik der Mathematik. *Math. phys. Semesterberichte, N.F.*, 21:164–169.

[Murier and Rousson, 1998] Murier, T. and Rousson, V. (1998). On the randomness of the decimals of π. *Student*, 2(3):237–246.

[Pollard, 1984] Pollard, D. (1984). *Convergence of stochastic processes*. Springer-Verlag, New York.

[Rissanen, 1987] Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society B*, 49(3):223–239 and 252–265.

[Schwarz, 1978] Schwarz, G. (1978). estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

[Tukey, 1993a] Tukey, J. W. (1993a). Discussion- Davies's data sets. Princeton University, Princeton.

[Tukey, 1993b] Tukey, J. W. (1993b). How Davies's data sets might reasonably be approached. Princeton University, Princeton.

[Tukey, 1993c] Tukey, J. W. (1993c). Issues relevant to an honest account of data-based inference, partially in the light of Laurie Davies's paper. Princeton University, Princeton.

[Vapnik, 1998] Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York.

50.1 Figures

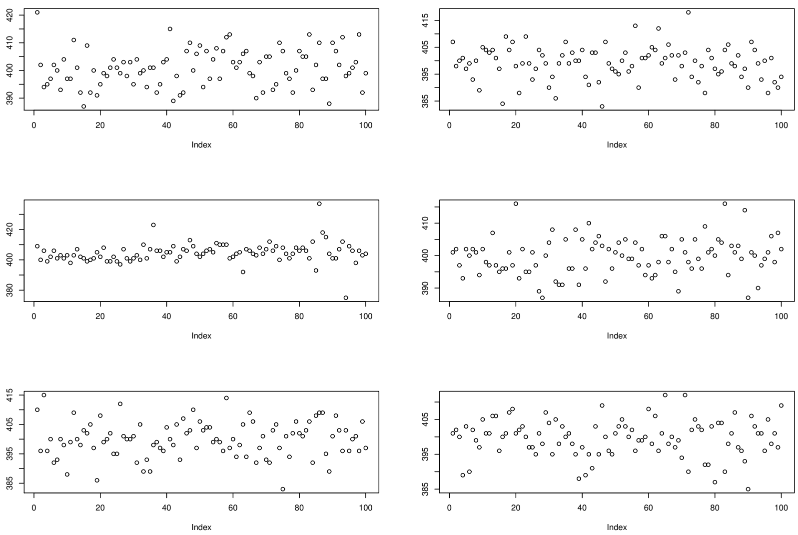

Figure 1:

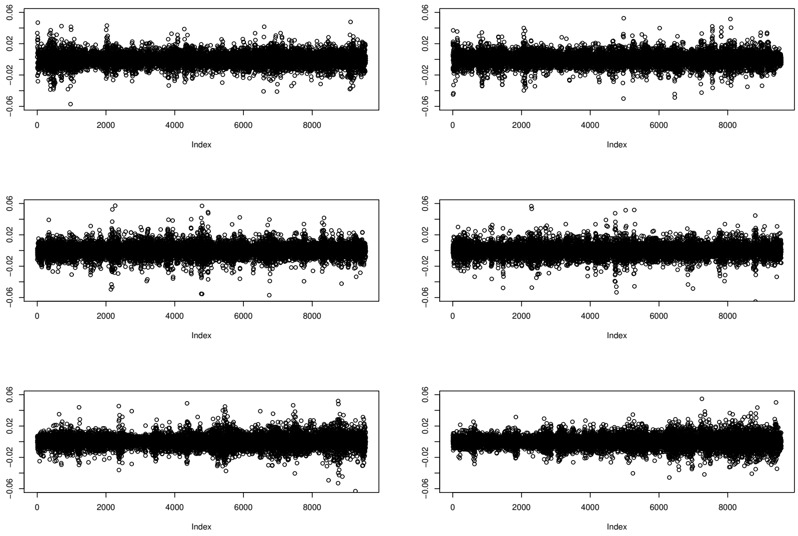

Figure 2:

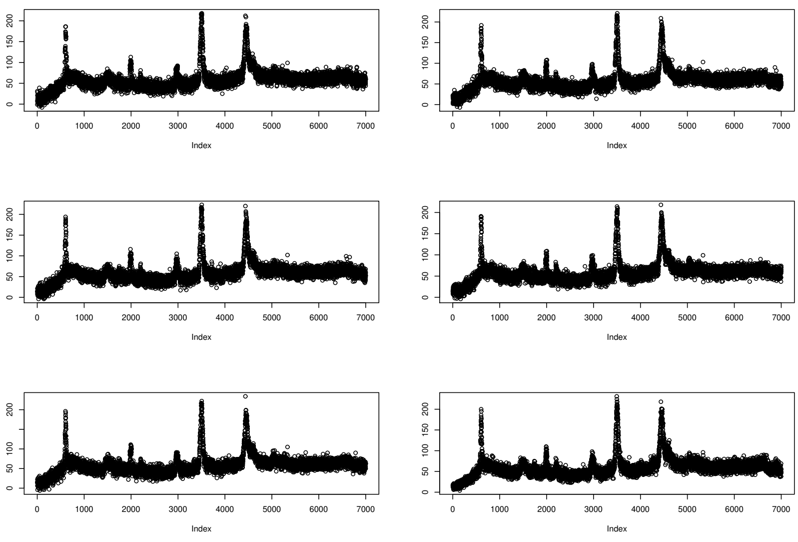

Figure 3:

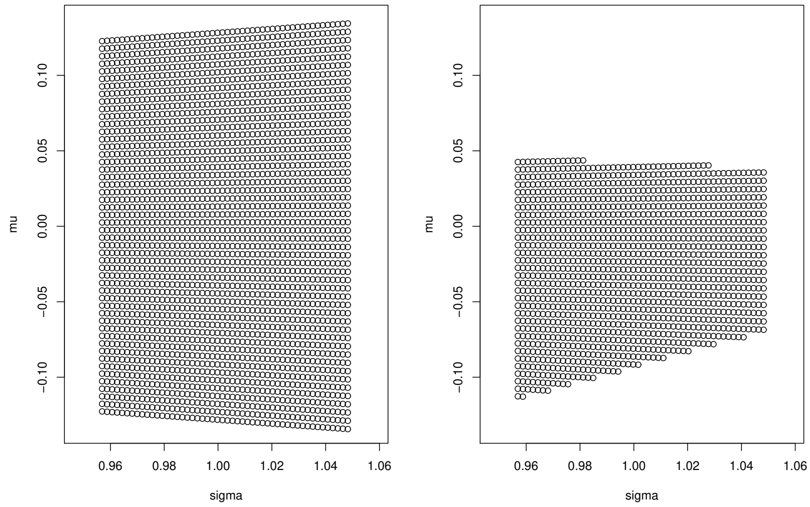

Figure 4:

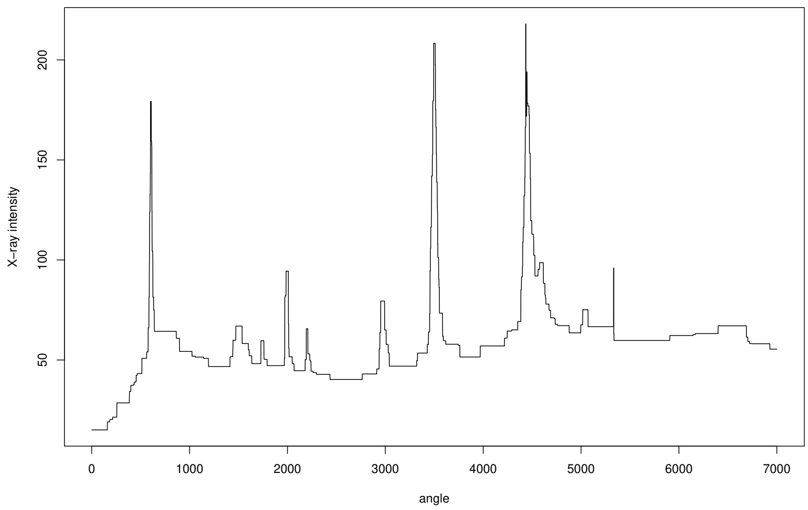

Figure 5: