# Approximating Data

P. L. Davies
*Fachbereich Mathematik und Informatik*
*Universität Essen*
*Essen, Germany*
*laurie.davies@uni-essen.de*

## 1. Approximation

The concept of approximation we use is a simple one. A probability model $P_\theta$ is an adequate approximation for the data set $(x_1, \ldots, x_n)$ if "typical" samples $(X_1(\theta), \ldots, X_n(\theta))$ of size $n$ generated using $P_\theta$ "look like" the real data set $(x_1, \ldots, x_n)$. The word "typical" is made precise by specifying a number $\alpha$, $0 < \alpha < 1$, such that $100\alpha\%$ of the generated samples "looks like" the real sample. The words "look like" are made precise by specifying some numerical feature or features which a data set may exhibit or not. To make the idea more concrete consider 999 data sets generated under the model $\theta$

$$\begin{array}{ccc} (X_{1,1}(\theta) & \ldots, & X_{1,n}(\theta)) \\ (X_{2,1}(\theta) & \ldots, & X_{2,n}(\theta)) \\ \vdots & \vdots & \vdots \\ (X_{999,1}(\theta) & \ldots, & X_{999,n}(\theta)) \end{array}$$

and insert the real data set at random. There are now in all 1000 data sets. The problem is to specify $1000(1-\alpha)$ of theses samples. If the real sample is not one of those specified then the model $\theta$ is regarded as an adequate approximation to the real data set. Figure 1 shows a small scale version of the idea. One of the six data sets is real and shows the result the spectroanalysis of the gall stone. The other five data sets are ones generated under a model. The problem is to identify the real data set. If this is not possible we define the model to be an adequate approximation. In practice the decision is made on the basis of one or more numbers calculated from each sample. We give an example.

**Example 1** (The multiresolution Gaussian noise approximation)
*Consider a sample $(x_1, \ldots, x_n)$ of size $n = 2^m$. The model we use is that of i.i.d. random variables with common distribution $N(0, \sigma^2)$. We put*

$$s_n = 1.483 * MED(|x_2 - x_1|, |x_3 - x_2|, \ldots, |x_n - x_{n-1}|)/\sqrt{2}$$

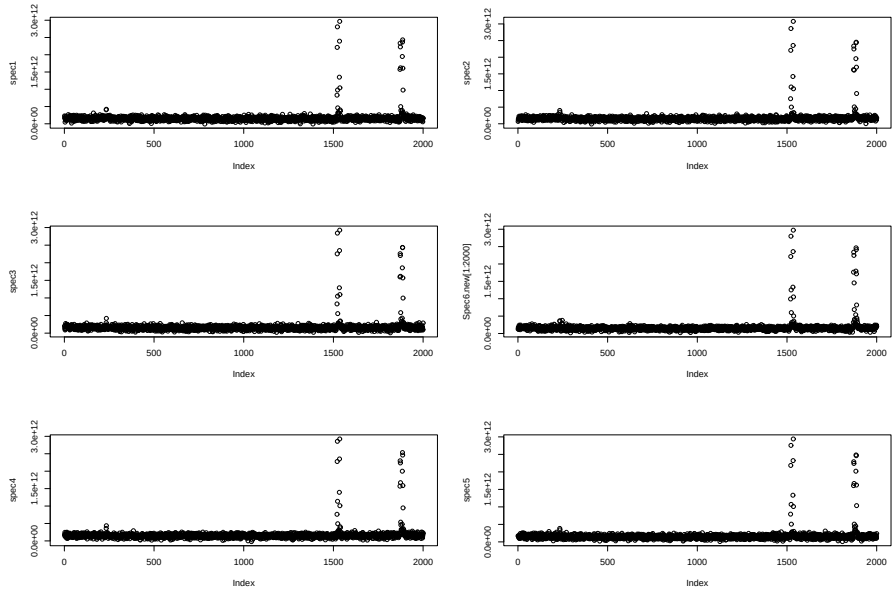


Figure 1: Five simulated and one real data set of the spectroscopic analysis of a gall bladder

*and define*

$$w_{ij} = 2^{-j/2} \sum_{l=(i-1)2^j+1}^{i2^j} x_l, \quad 1 \le i \le 2^{m-j}, 0 \le j \le m$$

*The model* $N(0, \sigma^2)$ *will be regarded as an adequate approximation to the data if*

*(a)* $|\sigma - s_n| \le 2.5\sigma/\sqrt{n}$

*(b)* $|w_{ij}| \le s_n\sqrt{2\log n}$ *for all* $i$ *and* $j$.

## 2. Topologies

The problem of finding the real data amongst a large number of simulated data generated according to some probability law would, on a theoretical level, often seem trivial. For the Gaussian model we simply choose that data set whose values are rationals. As all Gaussian random variables are rational with probability zero we would always be able to recognize the real data set. To avoid this we can agree to truncate the simulated data sets to the same degree of precision as the real data. Suppose the precision is given by $\epsilon$ and that the model under consideration is that of standard i.i.d. Gaussian random variables. Denoting the distribution function by $\Phi$ we consider another model, that of i.i.d. random variables with common

distribution function $F$. Suppose that $F$ satisfies

$$\sup_{0<p<1} |F^{-1}(p) - \Phi^{-1}(p)| < 0.001\epsilon. \tag{1}$$

We are able to generate coupled i.i.d samples $(X_1(\Phi), \ldots, X_n(\Phi))$ and $(X_1(F), \ldots, X_n(F))$ such that

$$\max_i |X_i(\Phi) - X_i(F)| < 0.001\epsilon.$$

From this it follows that the truncated samples will be equal with high probability. This implies that if $\Phi$ is an adequate model for a data set then so is $F$. Metrics similar to that of (1) are weak metrics with few open sets. Metrics of the form

$$d_{\mathcal{C}}(\mathbb{P}, \mathbb{Q}) = \sup_{C \in \mathcal{C}} |\mathbb{P}(C) - \mathbb{Q}(C)| \tag{2}$$

where $\mathcal{C}$ is a Vapnik-Cernovenkis class of sets are weak metrics. Weak metrics allow non-trivial direct comparisons between empirical measures and probability models. Any reasonable concept of approximation in statistics will, on a formal level, have to be continuous with respect to such weak topologies.

## 3. Models

There are no general principles which can be invoked to develop statistical procedures. By this we mean there is no "likelihood principle" and no Bayesian paradigm. Although procedures to not assume that the data follow a probability model we may derive a procedure from such a probability model.. Many data sets may be well approximated by Gaussian white noise. Consider a data set $y(t_i), i = 1, \ldots, n$ of values measured at times $t_i$. We look for a decomposition of the form

$$y(t_i) = f_n(t_i) + r_n(t_i), i = 1, \ldots, n \tag{3}$$

where $f_n$ is a simple function and the $r_n$ the resulting residuals. This corresponds to Tukey's decomposition

$$\text{DATA} = \text{SIGNAL} + \text{NOISE}. \tag{4}$$

We assume that the signal is simple and that the noise is complex. The simplicity of a function $f_n$ will be measured by the number of local extreme values. The complexity of the noise will be interpreted as being adequately approximated by Gaussian white noise. The definition of approximation we use is that of Example 1 above. This leads to the following problem of determining the minimum number of local extremes such that the residuals $r_n(t_i) = y(t_i) - f_n(t_i), i = 1, \ldots, n$ satisfy the approximation Example 1. Although a complete solution to the

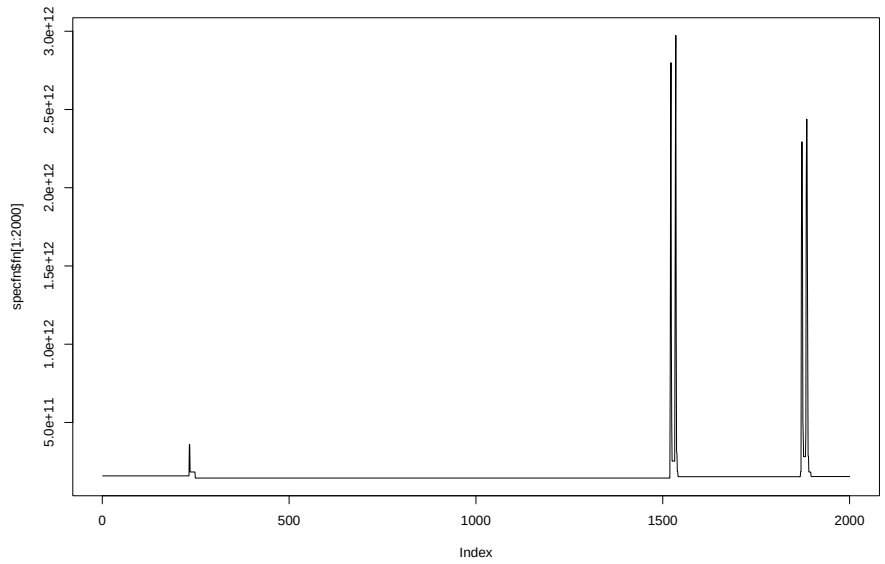


Figure 2: The regression function used in Figure 1

problem is not available Davies and Kovac (2001) provide a procedure which comes very close. Applied to the spectroscopy data of Figure 1 it gives the function shown in Figure 2. The real data set in Figure 1 is centre right. The others were simulated using the function shown in Figure 2 contaminated with Gaussian white noise with standard deviation given by the $s_n$ of Example 1 obtained from the original data. Although the procedure makes use of thresholding based on the Gaussian model it does not assume that the data really are Gaussian or that the noise is white. Indeed the noise in the spectroscopy data is slightly coloured but this has no major effect on the result.

## REFERENCES

Davies, P. L. (1995). Data features. *Statistica Neerlandica*, **49**, 185-245.
Davies, P. L. and Kovac, A. (2001) Local Extremes, Runs, Strings and Multiresolution. *Annals of Statistics*,to appear.

## RESUME

The paper develops a concept of approximation of data by probability models. It is argued that the topology involved should be weak. An example in the context of nonparametric regression is given.

Cet article expose une idé d'approximation des données par des modèles de probabilité. On propose que la topologie de l'analyse des donné doit être faible. Un exemple dans le contexte de la regression nonparametrique est donné.