Teaching Statistical Modeling through Smoothing

Joachim Engel
PH Ludwigsburg, Dept. of Mathematics and Computer Science
Reuteallee 46
71634 Ludwigsburg, Germany
engel_joachim@ph-ludwigsburg.de

On Modeling Scatterplot Data

Starting point for investigating functional relationships between two empirical variables are n pairs of measurements $(x_1, y_1), \ldots, (x_n, y_n)$ represented in a scatterplot. The objective of the modeling process is to derive a function f expressing the dependence of the two variables, either through a functional term of the form $y = f(x)$ or a function graph. A simple graph or functional term $y = f(x)$ representing the data cloud is an efficient compression of the data which is easy to communicate to others and easier to interpret and to compare with other graphs than the original data set. Important for gaining new insights into the observed empirical phenomena is to discern patterns and structures from the obtained graphical representation. The mathematically obtained result, if as functional term or as graph, may play a decisive role in the development of a theory for the empirical science the data originate from.

Interpolation - Curve Fitting - Smoothing

The simplest way to derive a graph from a scatterplot is through interpolation. Connecting the data with straight lines or by interpolating splines results in curves that may well help at discerning possible trends in the observations. Any type of interpolation is certainly appropriate when the observations represent error free measurements of the variables of interest, i.e. if $y_i = f(x_i)$ holds exactly. However, in most empirical situations the observations are subject to measurement errors, sampling errors or other disturbances ("noise"). This applies in particular to sampling models, i.e. when the observations are a random sample of a larger population of interest.

Recognizing that variables in empirical studies are usually disturbed by measurement and sampling errors leads to consider stochastic models. The most common approach is to decompose the observations additively into $y_i = f(x_i) + e_i$, where the model function f represents the trend and e_i is noise. The approach of curve fitting is based on the assumption that the function $f(x) = f(x, \vartheta)$ belongs to a pre-specified or known class of functions characterized by a finite dimensional parameter ϑ(e.g. linear, exponential, logistic function). Then the objective is to determine that value of the unknown parameter such that the model function fits the data best.

A problem with fitting curves from a parametric family is that their derivation may be guided by intuition and experience from the field of application, but it rarely has an objective justification. The parametric model just may be misspecified. This drawback calls for methods with more flexibility because the assumption of a parametric functional class becomes a "straight jacket" imposing a given structure on the data or excluding possibly existing data structure by assumptions - a contradiction to the principles of exploratory data analysis. Computationally intensive smoothing methods allow the derivation of model curves with a minimum of a-priori specifications. Explicitly or implicitly they are based on the concept of a weighted

moving average or other moving location parameter (for detailed discussion see, e.g., Wand and Jones, 1996 and Fan and Gijbels, 1995). Both methods are easy to motivate but both depend decisively on a parameter, the bandwidth. As a first approach in applications the bandwidth may be chosen experimentally and interactively with a PC. The computer then produces several graphical representations emphasizing various aspects in the data structure. Because of their flexibility nonparametric methods for modeling regression data are implemented in most statistics software packages. An experimental choice of the bandwidth leads to the questions of how much random noise the data entail, a question that is not addressed by interpolation nor by linear regression. The choice of bandwidth mediates between the two extremes of interpolation ($h = 0$) and linear regression line ($h = \infty$ for local linear approximation).

Systematic and Random Variation in Data

The techniques of curve fitting and moving average differ decisively from the interpolation approach: they leave room for measurement and sampling error and hence are the appropriate choice for modeling empirical data. The transition from the deterministic to the stochastic model is a qualitative step, which requires students to have an understanding of functions on a pre-calculus level as well as basic experience with randomness and chance. Has the concept of a function originally been introduced as a deterministic mapping and is probability in introductory courses often encountered as an investigation of "pure" randomness without a trend, these two concepts have to be considered together now. The discovery of trends in bivariate data–discerned first through visual inspection, then through numerical considerations and the use of modern technology–forms an important part of the data analysis curriculum (NCTM, 1998). When teaching about modeling scatterplot data I let my students first draw free-hand graphs, based on an eyeball inspection of the data before introducing scatterplot smoothers and curve fitting. Novices in probability and statistics tend to stick to a deterministic - mechanistic view of the world, which either doesn't allow room for chance or knows only trend free randomness. When considering noisy observations in empirical studies, the random part has to be separated from the deterministic trend. Here computer simulations offer the opportunity to develop and deepen a sense for random fluctuation in real data in order to focus on the relation between systematic and random in data.

REFERENCES

Fan J. and Gijbels, I. (1996): Local Polynomial Modelling and its Applications . Chapman and Hall: London.

National Council of Teachers of Mathematics (1998), Principles and Standards for School Mathematics: Standards 2000. Discussion Draft, Reston, Virginia

Wand, M.P. and Jones, M.C. (1995): Kernel Smoothing. Chapman and Hall: London.

FRENCH RÉSUMÉ

Le lissage de données bivariates est consideré comme une méthode entre interpolation et adaption d' un fonction paramétrique. Cet énoncé accentue la relation entre la variation systématique et la variation aléatoire en les données.