Data Smoothing

J. T. Grissom, and D. R. Koehler

Citation: American Journal of Physics **39**, 1314 (1971); doi: 10.1119/1.1976648
View online: https://doi.org/10.1119/1.1976648
View Table of Contents: http://aapt.scitation.org/toc/ajp/39/11
Published by the American Association of Physics Teachers

form $x^2=\eta t$. The classical problem of a rod of fixed length originally at a temperature T_0 and whose end faces are held at a temperature T cannot be solved by this technique.

We note that the transformation M is not completely determined due to the variability of n. Even with appropriate initial conditions for the transformed equation we have a one-parameter infinity of solutions to choose from. It is not clear that more than one physically meaningful solution should exist. This suggests the value of n that ultimately completely determines M should not be arbitrary but should reflect an important physical feature of the problem.

ACKNOWLEDGMENT

This work was supported by the Atomic Energy Commission.

[1] G. Birkhoff, *Hydrodynamics* (Princeton U. P. for the Univ. of Cincinnati, Princeton, N. J., 1950).
[2] A. G. Hansen, *Similarity Analysis of Boundary Value Problems in Engineering* (Prentice–Hall, Englewood Cliffs, N. J., 1964).
[3] W. F. Ames, *Nonlinear Partial Differential Equations in Engineering* (Academic, New York, 1965).
[4] A. G. Hansen and T. Y. Na, J. Basic Eng. **90, 71** (1968).
[5] A. J. A. Morgan, Quart. Appl. Math. **12,** 250 (1953)
[6] S. J. Kline, *Similitude and Approximation Theory* (McGraw–Hill, New York, 1965).

Data Smoothing

J. T. GRISSOM*
University of Tennessee
Knoxville, Tennessee 37916†
D. R. KOEHLER
Bulova Watch Company
Flushing, New York 11370
(Received 1 February 1971)

A class of methods is derived for data smoothing based on minimizing an error function consisting of two terms. The first term is the weighted sum of the squares of the deviations between the smoothed values and the original data, and the second term is the weighted sum of the squares of the $(k+1)$st order finite differences of the smoothed values. The method embodies two arbitrary parameters. Specifying the value of k chooses one member of the class of methods. The other parameter is a weighting factor that determines the degree of smoothing achieved at each point. Limits on the value of the weighting factor can be imposed based on the statistical properties of the data.

INTRODUCTION

Often the experimental physicist will encounter situations where it is desirable to make use of data smoothing techniques in the analysis or presentation of experimental results. An ideal example is the measurement of energy spectra with scintillation or solid state detectors using a multichannel pulse height analyzer. The resulting data is a digital representation of an intensity distribution as a function of energy. In addition to certain types of resolution distortion,[1] each data point can vary from the true value because of the presence of statistical fluctuations. It is the purpose of data smoothing to reduce the effects of such random variations in order that those features of the distribution that exceed the expected statistical deviations will become more discernible. Other obvious examples of the use of data smoothing techniques suggest themselves, but the underlying motivation is the same, namely, to reduce distortion resulting from statistical errors.

Perhaps the most commonly used smoothing procedure is based on the least squares fitting of polynomial functions to the data points.[2,3] Since the method demonstrates some of the basic features of data smoothing techniques in general, we will examine it in some detail. To illustrate

the procedure, consider a set of data $\{f_i\}$ that represents experimentally measured values of a function $f(x)$ at a discrete set of values $\{x_i\}$. We wish to replace the set of values $\{f_i\}$ by another set of values $\{F_i\}$ that represent the "smoothed" form of the data. To generate F_k, one chooses from $\{f_i\}$ the subset consisting of f_k and the m values on either side of f_k. Then an nth degree polynomial of the form

$$y=a_0+a_1x+a_2x^2+\cdots+a_nx^n$$

is fitted to these $2m+1$ points using a least squares criterion to generate the set of coefficients $\{a_n\}$. Once the coefficients have been determined from the least squares fit, F_k is just the value of the polynomial at x_k, i.e., $F_k=y(x_k)$. In this way the set $\{F_i\}$ is generated point by point. This procedure must, of course, be modified near the end points since for an arbitrary choice of m a value of f_k will not necessarily have m experimental values on both sides of it. This situation can be handled by always choosing the $2m+1$ points nearest the point being smoothed, with the additional requirement that the point being adjusted be the central value whenever the range of the data permits.

With various modifications the procedure outlined above forms the basis of most polynomial smoothing techniques. The method directly illustrates a basic assumption of any smoothing technique, namely, that the smoothed value of each point should be influenced in some manner by the values of the data points on either side of the point being adjusted. This is just another way of saying that the function for which the data are a statistically distorted representation is smooth, or continuous, and that the interval over which any rapid fluctuations in the value of the function occurs is large with respect to the interval between the measured values. Because of the resolution distortion characteristics of the experimental measuring apparatus this assumption is essentially always satisfied even for data which should in principle display discontinuities or very rapid fluctuations. The resolution of the measuring apparatus causes such discontinuities to be distorted into a smoothly varying function, and the proper design of the experiment dictates that measurements be performed in intervals that are on the order of, or smaller than, the "resolution width" of the measuring apparatus.[1]

Note that the polynomial smoothing technique is essentially a two-parameter method that is characteristic of any routine for smoothing a two-dimensional function for which the statistical uncertainty occurs in only one dimension. These parameters affect the resulting values of the smoothed data in a complex manner. First, m determines the number of points that exert any influence on the adjusted value of each data point. Once m is fixed, then n determines the degree to which each point is smoothed, i.e., the deviation between the smoothed value and the original value. For fixed m the degree of smoothing decreases with increasing n. For $n=0$ the smoothing is a maximum with the smoothed value of each point being just the average of the nearest $2m+1$ data points. For $n=2m$ no smoothing is accomplished since a polynomial of degree $2m$ will pass through all of $2m+1$ points. The range of n is thus $0\leq n\leq 2m$.

There are several aspects of this type of smoothing that are unsatisfactory from a mathematical point of view. The major difficulty is the absence of quantitative criteria on which to base a choice of m and n. It is difficult to defend even qualitatively a specific value for m. The only satisfactory operational criterion is that m must be sufficiently small that a polynomial of reasonable degree will give a good fit to the $2m+1$ points. This is a subjective criterion based on how rapidly the data fluctuate in the vicinity of the point being smoothed. A value of m that yields satisfactory results in one region of the data might conceivably be too large to work properly in another region. The choice of n is equally difficult to justify and in addition is related to the choice of m. Quite often the same degree of smoothing can be achieved by more than one set of values for m and n. What is even more difficult to justify *a priori* is the choice of polynomial functions to fit the data. There is no basis for choosing polynomials instead of a sum of exponentials or a sum of sines and cosines, for example. Here again such a choice is an intuitive or subjective one not based on any degree of rigor.

In the following analysis we present a more direct approach to the problem of data smoothing in which the significance of the arbitrary parame-

ters can be interpreted directly and for which quantitative criteria aid in choosing the parameters based on the statistical properties of the data.

I. A DIRECT APPROACH

Consider a set of n data points $\{f_i\}$ that represent the measured values of the function $f(x)$ with $f_i=f(x_i)$. Due to the presence of statistical variations in the measuring procedure each of the values f_i differs from the corresponding true value F_i with $F_i-f_i=\epsilon_i$. For the sake of simplicity we limit our attention to data where the error in the measurement of f_i greatly exceeds the expected error in x_i. Let σ_i denote the standard deviation in f_i. The value of σ_i is either experimentally determined or statistically estimated from the data.

Any smoothing technique must be based on a few fundamental assumptions. For a two-dimensional function $f(x)$, where the major statistical error is confined to only one dimension with relatively negligible error in the independent variable, the smoothing procedure should contain two arbitrary parameters as discussed above. One of these can be related to the number of neighboring points that influence each of the smoothed values, and the other parameter can determine the degree of smoothing at each point. These are the two degrees of freedom in the smoothing process. Note that the former degree of freedom is related to the independent variable while the latter one affects only the dependent variable. Consequently, the two-parameter nature of the smoothing process derives from the two-dimensional character of the data.

The smoothed data itself must satisfy two essential requirements. First, it must be a statistically valid representation of the original data, and second, it must represent a continuous function. A smoothing procedure should embody quantitative measures of the degree to which each of these requirements is satisfied. Quite obviously the two requirements cannot be totally independent but are in fact opposite conditions. The more the data are smoothed the less the agreement between the initial and smoothed values. For the sake of discussion we will refer to these two conditions as, respectively, the requirements of "reproducibility" and "smoothness."

Consider the quantity

$$E_1=\sum_{i=1}^{n}\frac{(F_i-f_i)^2}{\sigma_i^2}=\sum_{i=1}^{n}\frac{\epsilon_i^2}{\sigma_i^2}, \tag{1}$$

where E_1 is a positive quantity whose magnitude measures the agreement between the smoothed data and the original values. Each of the squared deviations in Eq. (1) is weighted by the inverse of the corresponding variance σ_i^2. Thus E_1 is a quantitative measure of reproducibility. Minimizing the value of E_1 with respect to the choice of $\{F_i\}$ in the least squares sense results in the trivial solution

$$E_1=0; \qquad F_i=f_i; \qquad i=1, 2, \cdots, n,$$

which satisfies the reproducibility criterion exactly but violates the smoothness requirement.

To measure the smoothness of the solution we will make use of the finite difference Taylor series representation

$$F_{i+s}=F_i+s\delta F_i+[s(s+1)/2!]\delta^2F_i+\cdots +[s(s+1)\cdots(s+k-1)/k!]\delta^kF_i+R_{ki},$$

where s is an integer, δ^kF_i represents the kth finite difference of F_i, and R_{ki} is the remainder after $k+1$ terms. We will limit ourselves to the case $s=1$ for which

$$F_{i+1}=F_i+\delta F_i+\delta^2F_i+\cdots+\delta^kF_i+R_{ki}. \tag{2}$$

The finite differences are of the form

$$\begin{aligned}
\delta F_i&=F_i-F_{i-1},\\
\delta^2F_i&=F_i-2F_{i-1}+F_{i-2},\\
\delta^3F_i&=F_i-3F_{i-1}+3F_{i-2}-F_{i-3},\\
&\vdots\\
\delta^kF_i&=\sum_{j=0}^{k}(-)^j\binom{k}{j}F_{i-j},
\end{aligned} \tag{3}$$

where

$$\binom{k}{j}=\frac{k!}{(k-j)!j!},$$

are the binomial coefficients. From Eqs. (2) and

(3) and the properties of the binomial coefficients one can show that

$$R_{ki}=F_{i+1}-\sum_{j=1}^{k+1}(-)^{j-1}\binom{k+1}{j}F_{i-(j-1)} \quad (4)$$

and

$$R_{ki}=\sum_{j=0}^{k+1}(-)^{j}\binom{k+1}{j}F_{i-(j-1)}=\delta^{k+1}F_{i+1}. \quad (5)$$

From Eq. (5) we see that the magnitudes of the R_{ki} directly indicate the relative smoothness of the solution $\{F_i\}$. Consider

$$E_2=\sum_{i=k+1}^{n-1}R_{ki}^2 \quad (6)$$

defined so that E_2 is a positive quantity whose value measures the smoothness of the $\{F_i\}$. Minimizing E_2 with respect to the choice of the $\{F_i\}$ in the least squares sense results in the trivial solution

$$E_2=0; \qquad F_i=\text{const}; \qquad i=1, 2, \cdots, n,$$

which satisfies the smoothness criterion exactly but in general violates the reproducibility requirement.

We can combine the error functions E_1 and E_2 that measure, respectively, the degree to which the reproducibility and smoothness criteria are individually satisfied into a single error function E such that

$$E=E_1+\alpha E_2, \quad (7)$$

where α is a constant. The value of E measures both the smoothness of the curve and the extent to which the smoothed values reproduce the initial data. The value of α determines the relative weight of E_2 with respect to E_1. Smoothing is accomplished by minimizing the value of E with respect to the choice of the $\{F_i\}$. The smoothed solution $\{F_i\}$ is thus determined by solving the set of n simultaneous equations

$$\partial E/\partial F_j=0; \qquad j=1, 2, \cdots, n, \quad (8)$$

with k and α fixed.

We have now formulated a least squares smoothing routine based on the arbitrary parameters k and α. The value of k specifies the number of terms in the finite difference series representation of Eq. (2), which is analogous at least to determining the number of neighboring points that are sampled to adjust the value of each of the initial data points. This analogy is not rigorous since in actuality the value of each F_i resulting from the solution of Eqs. (8) depends on all of the data values. The parameter α fixes the relative weights of E_1 and E_2 in the minimization of E, which is analogous to determining the degree of smoothing at each point. Later we will discuss the dependence of the solution on k and α.

Minimizing E by Eq. (8) automatically minimizes the deviations between the smoothed solution and the original values as far as is consistent with the degree of smoothing achieved. This in turn is determined by the value of k and α or more specifically, by the weighting factor α once k is specified. Thus neither E_1 nor E_2 is minimized absolutely, but their weighted sum is minimized to satisfy both the requirements of reproducibility and smoothness.

II. MATRIX FORMALISM

We will rewrite the smoothing procedure using a matrix formalism to make it easier to appreciate the mathematical structure of the technique and to interpret the significance of the various quantities that comprise the method.

Let $\mathbf{f}$ represent an $n\times 1$ column matrix whose elements consist of the n values $\{f_i\}$. Similarly $\mathbf{F}$ represents an $n\times 1$ matrix composed of the n smoothed values $\{F_i\}$. Then $\boldsymbol{\varepsilon}$ is the matrix given by

$$\boldsymbol{\varepsilon}=\mathbf{F}-\mathbf{f}. \quad (9)$$

If $\mathbf{S}$ is an $n\times n$ matrix whose elements s_{ij} are given by

$$s_{ij}=\sigma_i^2\delta_{ij}, \quad (10)$$

where δ_{ij} represents the usual Kronecker delta, then Eq. (1) can be written in matrix form as

$$E_1=(\mathbf{F}-\mathbf{f})^T\mathbf{S}^{-1}(\mathbf{F}-\mathbf{f}) \quad (11a)$$

or

$$E_1=\boldsymbol{\varepsilon}^T\mathbf{S}^{-1}\boldsymbol{\varepsilon} \quad (11b)$$

in terms of the matrix $\boldsymbol{\varepsilon}$ and its transpose.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.