# Mathematical Evaluation of the Fit of a Theory with Experimental Data

Otto Exner*

Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, 166 10 Prague 6, Czech Republic

Ivan Kramosil

Institute of Computer and Information Science, Czech Academy of Sciences, 182 07 Prague 8, Czech Republic

Igor Vajda*

Institute of Information Theory and Automation, Czech Academy of Sciences, 182 08 Prague 8, Czech Republic

Received July 15, 1992

A statistical evaluation of the fit of a theoretical conception (model) with experimental data has been almost always based on the assumption that experiments are loaded with errors while theory is simply either right or wrong. In this paper an alternative model is suggested, assuming that the experiments are effectively exact as compared to the approximate theoretical description. Basic problems of this model are formulated. For the simplest problem, comparison of two theories on the basis of one data set, two general solutions are suggested: one using nonparametric statistics; one syntactical in character.

## 1. INTRODUCTION

Comparison of theoretical and experimental results is of importance from both sides, i.e. with the intention of either checking the theory or interpreting the experiments. Further discussion will be restricted to experiments carried out on separate independent objects and does not concern quantities which are functions of any other independent parameter (e.g. time, temperature). It follows that our data are represented by an unordered set of experimental quantities $X_i$: the corresponding theoretical quantities $Y_i$ (each $Y_i$ belongs to the pertinent $X_i$) are assumed to be obtainable by reproducible calculations. For comparison of $X_i$ with $Y_i$ two mathematical frameworks are well known:

(i) In classical statistics, hypothesis testing,[1] $X_i$ are assumed to be loaded with random errors $\epsilon_i$, whose distribution can be specified. The theory is not connected with any error; it is either right or wrong. Assuming that it is right, the likelihood of $X_i$ can be obtained. If this likelihood is lower than a given value, the theory is rejected. Otherwise the theory is acceptable within the assumed experimental error. The statistical induction leads in this way from the experiments actually carried out to all experiments possible.

(ii) In the calculus of observations the experiments are loaded with error as above, but the theory is in principle incontestable. (For instance the sum of the angles in a triangle is 180°.) The task is to modify the $X_i$ values in order to get a perfect agreement.

A need for developing the third approach is apparent in certain sciences which are not quite exact but not purely descriptive in character. In our opinion typical is chemistry[2] which has at its disposal a lot of numerical data (compared, e.g., to botany), but even the sophisticated theories are only approximate (compared to physics): in such cases one speaks merely about models instead of theories. As the experiments are gradually improved, approach i could lead to rejecting all the models actually considered without giving a possibility of classifying them; i.e. which ones are acceptable, or which one is better than another. For this reason, just in chemistry attempts have appeared to define statistics measuring the fit.[3-5] It is true that the opinions were offered many times that even the model need not be quite exact; however, procedure i was not changed. In many applications (e.g., refs 4 and 5) it is not clear whether the difference should be attributed to the experimental errors or to the imperfection of the theory. For this reason we propose here the third, complementary approach:

(iii) Contrary to case i, experiments can be considered to be free of error; the difference $\Delta_i = X_i - Y_i$ will serve to evaluate a theory or to compare several theories to each other. The values $\Delta_i$ could be viewed as random variables if the objects were treated as if they were sampled at random from the universe of objects and if the relative frequencies of events $\Delta_1 \in E, ..., \Delta_n \in E$ were asymptotically (for $n \to \infty$) stable for reasonable sets $E$ (with limits satisfying the well-known postulates of probability theory). This assumption is quite restrictive in the given context, and not easy to verify. In other terms, we are unable to define any universe from which the objects would be sampled and we cannot secure that the selection was random.

Moreover, one can hardly imagine the procedures of specifying the distribution of the random vector $(\Delta_1, ..., \Delta_n)$. One could perhaps apply in this area the fuzzy set theories, having less strict postulates than those of probability theory,[6] but little is known so far about the methods of inference based on such theories.

## 2. STATEMENT OF THE PROBLEM

In this section we present several typical problems which may arise; later on we suggest two possible solutions of the first problem.

(a) The experiments are free of error: this is of course the limiting theoretical case. In practice, it will be sufficient when the experiments are much more exact than is the observed disagreement with the theory. What is important in this case is that the theory is independent from the experiments under consideration: if it uses some experimental quantities, they

0095-2338/93/1633-0407$04.00/0 

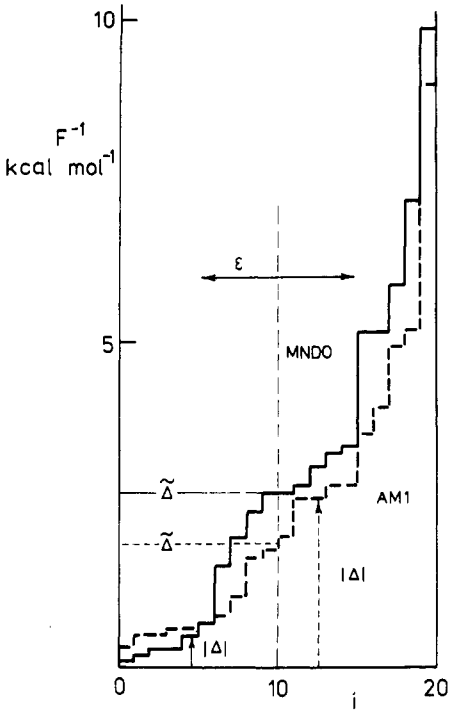


**Figure 1.** Evaluation of two quantum chemical theories, AM1 (full lines) and MNDO (broken lines), according to our statistical approach, eq 4, on the basis of 20 data from ref 10. According to eq 3, $F^{-1}$ is plotted versus the order of the experiment $i$. Theory AM1 is preferred in the range $\epsilon$, the reliability of the decision $\rho = 0.4$ according to eq 5.

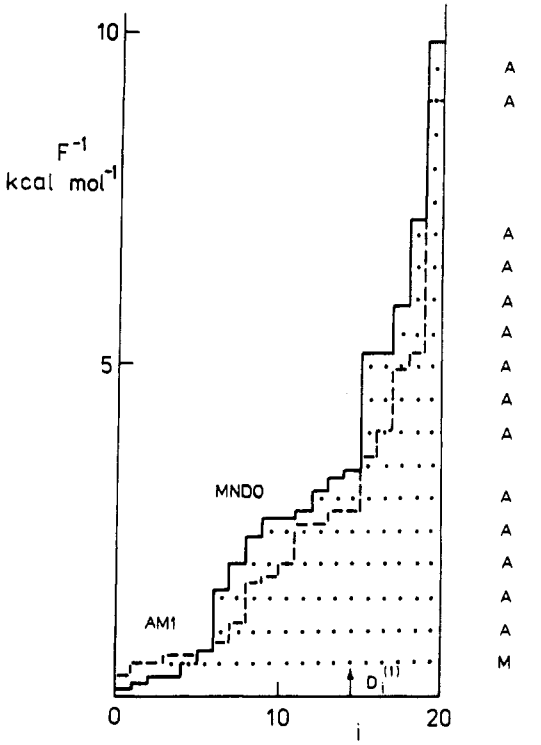


**Figure 2.** Evaluation of the same theories as in Figure 1 and on the basis of the same data, according to our syntactical approach, eq 8. The plot of $F^{-1}$ is the same as in Figure 1. Decision is made in each row according to the number of points right from the lines: theory with a smaller number of points is preferred, and the result is expressed by letters A and M, no letter means undecided. Theory AM1 is strongly preferred.

are related to previous exact experiments (like Planck constant, velocity of light). The task is to decide which of two theories, $Y_i$ or $Z_i$, reproduces better the given set $X_i$.

(b) A more general task than in the preceding case is to deal with two theories, $Y$ and $Z$, each relating to another set of experimental data, $X$ and $U$, respectively. In a particular case, the two latter sets may represent two properties of the same set of objects. The task is to decide whether there is a better agreement between $Y$ and $Z$, or between $X$ and $U$. When a number of data is accumulated, one could attempt to specify conventional values of the characteristics, corresponding to "excellent", "good", or "insufficient" theories (see, for example, the conventional scales for the correlation coefficient[7] and the criticism[3]).

(c) The theory is not independent of the pertinent experiments but has been constructed just to fit the given data set: most frequently it contains several parameters which have been optimized. The task in this case is to compare two theories as above. Typically one of them gives a better fit at the cost of more parameters. In an extreme case the underlying model may be very simple (e.g., a linear equation); on the other hand the parameters are numerous: this is the case of models obtained, e.g., from the principal component analysis.[8]

(d) The experiments are not perfectly accurate but loaded with a certain error. This error is generally smaller than $\Delta$, and its distribution can be specified in advance. The problems are as described above.

Evidently the problem denoted by (a) is the simplest. It is also the fundamental one since the others can be derived from it by releasing some constraints. We propose here two approaches for comparing two theories, $Y = (Y_1, ..., Y_N)$ and $Z = (Z_1, ..., Z_N)$ on the experimental data $X = (X_1, ..., X_N)$: they are called the statistical and the syntactical approach. We will now explain shortly the mathematical background. Even without any mathematics, our procedure is intelligible from the example and from Figures 1 and 2, which should be self-explanatory.

## 3. STATISTICAL APPROACH

The motivation of this approach is easier if we treat $(\Delta_1, ..., \Delta_N)$ as a vector of random variables. If there were a ground to assume that this vector is normal with zero mean, then the maximum likelihood decision in favor of the theory $Y$ would be given in terms of least squares of errors (observed variances)

$$\frac{1}{N}\sum_{i=1}^{N}[X_i - Y_i]^2 < \frac{1}{N}\sum_{i=1}^{N}[X_i - Z_i]^2$$

As discussed in the Introduction, the normality assumption is unfounded here, and, as is well-known from the robust statistics,[9] the good properties of the least squares inference are sharply deteriorating as soon as we depart even slightly from the pure normality. The methods based on medians $\text{med}\{|\Delta_i|: i = 1, ..., N\}$ of absolute deviations $|\Delta_1|, ..., |\Delta_n|$ are more robust with respect to possible misspecification of the true distribution of $(\Delta_1, ..., \Delta_N)$ in the sense exactly stated in ref 9. The decision in favor of theory $Y$ based on medians of absolute deviations takes place if

$$\text{med}\{|X_i - Y_i|: i = 1, ..., N\} < \text{med}\{|X_i - Z_i|: i = 1, ..., N\} \quad (1)$$

The number

$$\rho = \sup \epsilon$$

where the supremum extends over $\epsilon \in [0, 1]$ such that all $\alpha$-quantiles $\alpha\{|\Delta_i|: i = 1, ..., N\}$, $\alpha \in [(1 - \epsilon)/2, (1 + \epsilon)/2]$,

satisfy the relation

$$\alpha\{|X_i - Y_i|:\ i = 1, ..., N\} \leq \alpha\{|X_i - Z_i|:\ i = 1, ..., N\}$$

characterizes the significance of this decision.

This is the method which we suggest for use also in the case considered in the present paper, where there is no ground to assume that $(\Delta_1, ..., \Delta_N)$ is a vector of random variables. We now describe this method in a more formal manner.

For $Y$ and $Z$ the functions of real variable u are defined as follows:

$$F_Y(u) = \text{card}\{1 \leq i \leq N:\ |Y_i - X_i| < u\}$$

$$F_Z(u) = \text{card}\{1 \leq i \leq N:\ |Z_i - X_i| < u\} \qquad (2)$$

where card stands for the cardinality of the set in question, i.e., for the number of its elements in the case of finite sets. These functions can be interpreted as empirical distribution functions of absolute deviations of the first and second theory from the experiment. By $F_Y^{-1}(\alpha)$, $F_Z^{-1}(\alpha)$ we denote the corresponding right-continuous "quantile functions" in the domain $0 \leq \alpha \leq N$, e.g.,

$$F_Y^{-1}(\alpha) = \inf\{u > 0:\ F_Y(u) \leq \alpha\} \qquad (3)$$

and we consider the sets $\mathcal{E}_{Y,Z}$, $\mathcal{E}_{Z,Y}$ defined by

$$\mathcal{E}_{Y,Z} = \{\epsilon \in [0, N]:\ F_Y^{-1}(\alpha) \leq F_Z^{-1}(\alpha) \text{ for all } \alpha \in [(N-\epsilon)/2, (N+\epsilon)/2]\}$$

Obviously, at least one of the sets $\mathcal{E}_{Y,Z}$, $\mathcal{E}_{Z,Y}$ is nonempty.

The decision $T$ about which theory is better is defined by

$$T(X,Y,Z) = \begin{cases} Y & \text{if } F_Y^{-1}\left(\frac{N}{2}\right) < F_Z^{-1}\left(\frac{N}{2}\right) \\ 0 & \text{if } F_Y^{-1}\left(\frac{N}{2}\right) = F_Z^{-1}\left(\frac{N}{2}\right) \\ Z & \text{if } F_Z^{-1}\left(\frac{N}{2}\right) < F_Y^{-1}\left(\frac{N}{2}\right) \end{cases} \qquad (4)$$

Obviously, the rule eq 4 is equivalent with eq 1. If $T(X, Y, Z) = 0$, then no preference between the theories is made. If $T(X, Y, Z) = Y$, then

$$\rho = \frac{1}{N}\sup\mathcal{E}_{Y,Z} \in [0, 1] \qquad (5)$$

is called a reliability of the decision. The decisions with reliabilities close to 1 are safe. Otherwise the decisions have to be used with a caution proportional to the nonreliability $1 - \rho$.

**Example.** We have chosen an example from quantum chemistry which is a typical field of application of our procedures. The purpose is to illustrate the practical application, not to demonstrate the efficiency of the method. The heats of formation of a set of rather diverse compounds were calculated[10] by two semiempirical methods, the new AM1 and the standard MNDO, and compared with the experimental values. For the present purpose we shall consider these experimental values as free of error. The number of data was restricted to $N = 20$, to obtain a more lucid graphical representation (Figure 1). Figure 1 was constructed as follows. Differences $\Delta_i$ of experimental and theoretical values were calculated for 20 compounds using the method MNDO. Their absolute values $|\Delta_i|$ were ordered according to their magnitude and plotted against the serial numbers $i = 1, 2, ..., 20$ (full line). The procedure was repeated for the method AM1 (dashed line). The two lines represent our functions $F_Y^{-1}(\alpha)$ and $F_Z^{-1}(\alpha)$, respectively. In the middle of the diagram, at $i = 10$, the value for AM1 is lower; hence this method is given preference. The range $\epsilon$, in which AM1 is still preferable, is defined in such a way that it must be symmetrical around the middle of the graph: in our case it is restricted by the crossing in the region of low values, at $i = 6$. The reliability value $\rho$ is the ratio of the range $\epsilon$ to the whole breadth of the graph. It is seen that the reliability of the established preference is $\rho = 0.4$, which is quite far from 1.

We can still examine a larger set of 58 hydrocarbons which is the largest actually homogeneous sample obtainable from the given source.[10] Now MNDO is preferred only with $\rho = 0.1$. In the same sample AM1 appeared slightly better according to the conventional mean square error.[10] As expected the results may depend rather sensitively on the sample, particularly when the samples are small and not sufficiently homogeneous.

## 4. SYNTACTICAL APPROACH

This approach, syntactical or "nominalistic" in character, considers the values of $X, Y, Z$ as individual syntactical objects without reference to any model of their origin. The intuitive question as to which of the values $(Y, Z)$ is a "better" approximation may be formulated more precisely in the following way. For each $i \leq N$ let us define a sequence of functions $D_i^{(1)}, D_i^{(2)}, D_i^{(3)}, ...$, of an argument $D$ such that (1) for each $i$ fixed $D_i^{(j)} > 0$ and tends to infinity for $j \to \infty$ and (2) for each $i, j$ fixed $D_i^{(j)}(D) \to 0$ for $D \to 0$. Some possible, particularly simple functions are $D_i^{(j)} = jD$ or $D_i^{(j)} = q^jD$. The intuition behind is as follows. For each examined item, i.e., for each $i \in N$, $D_i^{(j)}$, $j = 1, 2, ...$, is an increasing and hence, a more and more tolerant sequence of intervals inside of which the difference $|X_i - Y_i|$ is taken as negligible; in other words, a prediction $Y_i$ is taken as acceptable for the actual value $X_i$. So, for a $j$ large enough, the prediction $Y_i$ will be always acceptable and we shall be interested in the maximum $j$ for which this is still not the case. The value of $D$ does not need any extramathematical interpretation. It is just a free parameter which enables one to define explicitly systems of tolerance intervals with arbitrarily narrow intervals of the lowest indices.

Denote $D^{(n)} = (D_1^{(n)}, D_2^{(n)}, ..., D_N^{(n)})$ and set

$$D[X, Y, D^{(n)}] = \text{card}\{i:\ i \leq N, |X_i - Y_i| > D_i^{(n)}\} \qquad (6)$$

or, more generally,

$$d[X, Y, D^{(n)}] = \text{card}\{i:\ i \leq N, |X_i - Y_i| > D_i^{(n)-}\} + \text{card}\{i:\ i \leq N, |Y_i - X_i| > D_i^{(n)+}\} \qquad (7)$$

supposing we consider two sequences $\{D_i^{(n)-}\}_{i=1,n=1}^{N,\infty}$ and $\{D_i^{(n)+}\}_{i=1,n=1}^{N,\infty}$ of appropriate positive reals. The interpretation for this case is an immediate generalization of the more simple one-sided case explained above. Evidently, the value $d(X, Y, D^{(n)})$ expresses the number of examined items for which $Y_i$ is an appropriate prediction for $X_i$ at the $n$th level of tolerance intervals, and it is independent of a permutation of indices 1, 2, ..., $N$.

The following decision function, $T_0$, is defined:

$$T_0[X, Y, Z, D^{(n)}] = \begin{cases} Y, & \text{if } d[X, Y, D^{(n)}] < d[X, Z, D^{(n)}] \\ 0, & \text{if } d[X, Y, D^{(n)}] = d[X, Z, D^{(n)}] \\ Z, & \text{if } d[X, Y, D^{(n)}] > d[X, Z, D^{(n)}] \end{cases} \qquad (8)$$

For $\bar{D} = \{D_i^{(n)}\}_{i=1,n=1}^{N,\infty}$, the "global" decision function $T^+$ reads

$$T^+(X, Y, Z, \bar{D}) = Y, \text{ if } \mathrm{card}\{n: T_0(X, Y, Z, D^{(n)}) = Y\} > \mathrm{card}\{n: T_0(X, Y, Z, D^{(n)}) = Z\}$$

$$= 0, \text{ when those cardinalities are equal}$$

$$= Z, \text{ in the case of the inverse inequality} \quad (9)$$

The interpretation of these three decisions is self-evident: $T^+(X, Y, Z, \bar{D}) = Y$ means that $Y$ is a better approximation of $X$ than $Z$ with respect to $D$, analogously for $T^+(X, Y, Z, \bar{D}) = Z$; when $T^+(X, Y, Z, \bar{D}) = 0$, then we are not able to decide. As $D_i^{(n)} \to \infty$ for $n \to \infty$ and for each $i \leq N$, there exists $n_0$ such that $T_0(X, Y, Z, D^{(n)}) = 0$ for each $n \geq n_0$; hence, both the sets $\{n: T_0(X, Y, Z, D^{(n)}) = Y\}$ and $\{n: T_0(X, Y, Z, D^{(n)}) = Z\}$ are finite.

This approach can be extended even to problem d, where the observations $X_1, X_2, \ldots, X_N$ are charged with an error of statistical kind. In this case we may replace each $X_i$ by an interval $\langle X_i^1, X_i^2 \rangle$, e.g., by a tolerance interval containing the "true" value with a probability exceeding an a priori given threshold value. The advantage of the syntactical method is two-fold. First, it does not depend on any assumptions concerning some model or mechanism by which the values $X_i$ and $Y_i$ have been generated or obtained; it takes and treats them just like sequences of real numbers without any interpretation, hence, syntactically. Second, the system $\{D_i^{(j)}(D)\}$ of threshold values seems to be flexible enough to be able to reflect even rather fine detailed a priori knowledge owned by the decision-making subject and concerning the specific features of the area under investigation.

**Example.** The approach was applied to the same data as those used previously for the set with $N = 20$. The beginning is the same as in the statistical approach. One obtains the same two plots of the absolute values $|\Delta_i|$ vs $i$ (Figure 2): the full line is pertinent to the method MNDO; the dashed line, to the method AM1. The next task is to choose a function $D_i^{(j)}$ suitable for the given problem. The simplest function $D_i^{(j)} = jD$, with $D = 0.5$ kcal mol$^{-1}$ was chosen here only for the sake of its simplicity: we do not claim that it should be the best choice in our case. Values of the argument $D$ can be represented as a series of heavy points forming something like a raster. Decisions are made according to the numbers of points right from the respective line. Where this number is smaller, the respective method is given preference for the given $j$ in $D_i^{(j)}$. The decisions are shown by the letters $A$ and $M$ in each row: the AM1 theory is preferred in 14 cases, MNDO in 1 case, and 4 cases are undecided. The result is insensitive to small modifications of the given choice of $D$.

For the sample of 58 hydrocarbons with $D = 0.5$, AM1 is preferred in 71 cases, MNDO in 10, 18 undecided. However, the results are quite sensitive to the form of the function $D_i^{(j)}$. By a proper choice of this function the whole approach can be modified quite dramatically. This will be discussed in the next section.

## 5. DISCUSSION

The two approaches as outlined above differ primarily in their sensitivity to outlying observations. (According to our definition, the term outlier can mean only failure of the theoretical treatment in an atypical case, not a big experimental error.) With respect to these outliers our statistical approach is intentionally robust. Its results are in fact independent of the extreme values of $\Delta$: compared to the customary criterion of mean square error, they depend always more on the middle part of the distribution curve. The syntactical approach is much more flexible through the proper choice of the function $D_i^{(j)}$. With a linear function ($D_i^{(j)} = jD$) the decision is very sensitive to extreme values which may become fully responsible for the result: then the decision is determined by the intention to avoid large errors (the minimax philosophy). By means of an exponential function ($D_i^{(j)} = q^jD$), the effect of extreme values may be deliberately suppressed according to the value of $q$. Then the decision may become either more or less sensitive to the outliers than is the customary mean square error. Our objection against the latter characteristic is mainly that it is neither robust nor flexible. It should not be used for the problems outlined here unless it has been proven that the distribution of $\Delta$ is normal. We have encountered some examples with distributions quite far from the normal ones.

Both the statistical approach and the syntactical approach with $D_i^{(j)}$ independent of $i$ are based on the underlying assumption that a certain value of the error $\Delta$ is of the same consequence, whatever is the value of $X_i$ itself. Quantitites satisfying this condition were called intensive;[2] they are often defined on a given (conventional) scale. (For instance in measuring temperature in °C, the agreement between 10 and 11 °C is of the same precision as that between 1 and 2 °C.) On the other hand, the extensive quantities are defined as multiples of a measuring unit and are proportional to the amount of matter. (For instance in measuring spectral intensity the agreement between 10 and 11 is the same as that between 1 and 1.1.) In this case error $\Delta$ should be related to the value of $X$. This can be achieved in the syntactical approach by incorporating $X$ into the function $D_i^{(j)}$, e.g., $D_i^{(j)} = jdX_i$, in the statistical approach, e.g., by transforming $X$ into logarithm.

All the considerations hitherto have used only absolute values of the differences $\Delta$, assuming tacitly that positive and negative errors are of the same consequence. This assumption can be related to the general postulate that the theory has no systematic error: it means that the expected value of $\Delta$ is zero. This may hold very generally for empirical theories, the problem b in section 2. For nonempirical theories this need not be true, and a simple test would be desirable in each case.

## 6. CONCLUSIONS

We believe that the above suggestions may be useful in comparing experiments with theories in the simplest case. Extensions to more complex cases, given b–d may be more or less evident; e.g., the syntactical approach can be easily extended to $X$ loaded with an experimental error (case d). We hope to proceed later to these questions. The intention of this paper was mainly to draw attention to the basic problems which in our opinion have received insufficient attention, compared to the great effort given to both experiments and theories.

## ACKNOWLEDGMENT

We acknowledge valuable discussions with Professor S. Wold during the first stage of our investigations when the first ideas were formulated.

## REFERENCES AND NOTES

(1) Rao, C. R. *Linear Statistical Inference and Its Applications*; Wiley: New York, 1965.
(2) Exner, O. *Correlation Analysis of Chemical Data*; Plenum Press: New York, 1988.

(3) Exner, O. Additive Physical Properties. I. General Relationships and Problems of Statistical Nature. *Collect. Czech. Chem. Commun.* **1966**, *31*, 3222–3251.
(4) Hamilton, W. C. Significance Tests on the Crystallographic *R* Factor. *Acta Crystallogr.* **1965**, *18*, 502–510.
(5) Ehrenson, S. On the *f* Statistic and Comparable Measures in Linear Free-Energy Relationship Fittings. *J. Org. Chem.* **1979**, *44*, 1793–1797.
(6) Dubois, D.; Prade, H. *Fuzzy Sets and Systems: Theory and Applications*; Academic Press: New York, 1980.
(7) Jaffé, H. H. A Reexamination of the Hammett Equation. *Chem. Rev.* **1953**, *53*, 191–261.
(8) Wold, S.; Sjöström, M. In *Chemometrics: Theory and Applications*; Kowalski, B. R., Ed.; ACS Symposium Series 52; American Chemical Society: Washington, D.C., 1977; p 243.
(9) Huber, P. J. *Robust Statistics*; Wiley: New York, 1981.
(10) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.