Model Discrimination and Criticism with Single-Response Data

Warren E. Stewart and Thomas L. Henson
Dept. of Chemical Engineering, University of Wisconsin, Madison, WI 53706

George E. P. Box
Center for Quality and Productivity Improvement, University of Wisconsin, Madison, WI 53706

The inverse probability theorem of Bayes is used, along with sampling theory, to obtain objective criteria for choosing among rival models. Formulas are given for the relative posterior probabilities of candidate models and for their goodness of fit, when the models are fitted to a common data set with Normally distributed errors. Cases of full, partial and minimal variance information are treated. The formulas are demonstrated with three examples, including a kinetic study of a catalytic reaction.

Introduction

It is helpful, in discussions of process modeling, to distinguish between empirical and mechanistic models. Consider first what we might mean by a "true" mechanistic model. Suppose that a measured response or output $y$, such as the yield of a particular product in a chemical process, was known to depend upon certain input variables $\xi_1, \ldots, \xi_k$ such as initial reactant concentrations, temperature, and pressure. Because of experimental errors, the output $y$ in replicate trials would fluctuate around a typical value called the mathematical expectation $E(y)$. This quantity is the mean value of $y$ over many conceptual repetitions of the experiment with the same settings of the input variables.

Suppose that a model is available that embodies the physical mechanism of the experimental system, so that the expectation of $y$ at each value $\xi$ of the experimental conditions is given exactly by

$$E(y) = f(\xi, \theta),$$  \hspace{1cm} (1)

where $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$ is a vector of fundamental parameters such as activation energies or diffusion coefficients. Then we shall say that Eq. 1 is a true mechanistic model of the measured phenomenon. It is not implied that for any given case such a functional form $f(\xi, \theta)$ is known, or even that it is knowable. A true model is, strictly speaking, a hypothetical concept that arises from our faith that physical phenomena ought to be explicable in mechanistic terms. Furthermore, although in some cases such a model might be expressible explicitly in terms of known functions, more often it would be definable only in terms of differential or integral equations. The methods in this article are applicable to such models when implemented with modern equation-solving methods.

We now consider what might be meant by an empirical model. Over limited regions of experimental conditions $\xi$ it would often be true that the relationship between $E(y)$ and $\xi$ was smooth and could be locally approximated by an interpolation function, $g(\xi, \theta)$. Then $g(\xi, \theta)$ might be used over such regions as a mathematical French curve to represent $E(y)$. For example, multidimensional polynomials have been used successfully for such empirical representation over limited ranges (Box and Wilson, 1951; Box, 1954; Box and Youle, 1955; Box and Hunter, 1957; Hill and Hunter, 1966).

Now the true mechanistic model and the purely empirical model represent extremes. The former would be appropriate in the extreme case where the mechanism was fully known, and the latter in the opposite extreme where the knowledge consisted only of the observations and some smoothness assumptions. The situation in most real investigations is somewhere in between, and as experimentation and learning proceed, the models used may show more understanding of mechanism (Box and Youle, 1955). Since real problems may occur anywhere between the two extremes, various statistical tools are needed to cope with them.

In some instances where almost nothing is known or accessible about the mechanism, only a rough local mapping of the response may be obtainable. Such rough mappings, neverthe-
less, may have great value. In cases where even a little more is known, careful thought can often lead to empirical models that reflect important known features of the system. For example, without detailed knowledge of the mechanism we may nevertheless know that the function \( E(y) \) must approach an asymptote for large values of some variable \( \xi \). We can then represent \( E(y) \) by a function that has such a form rather than, say, a second-degree polynomial (Box and Cox, 1964).

Other situations occur where the hope of representing the major features of \( E(y) \) by a mechanistic model is a reasonable one. We might wish to obtain such a model for either or both of the following reasons:

1. Basic intellectual curiosity might urge us to find out what was happening in a particular system, and such understanding may have great value. In cases where even a little more is known, careful thought can often lead to empirical models.

2. We might hope to develop a model that permitted some extrapolation, at least to indicate regions of the space where further investigation could be useful. Extrapolation is risky with any model, but becomes more meaningful as the model comes closer to the actual mechanism.

In the present article we address only one special aspect of the wide class of problems just implied—that of using existing data to discriminate among two or more candidate mechanistic models. The other problems mentioned earlier are of equal importance and have been discussed elsewhere (Box and Couttie, 1956; Box and Lucas, 1959; Box, 1960; Box and Hunter, 1962, 1965). In particular, a very important problem is that of choosing experimental conditions that will best discriminate among a set of mechanistic models. The latter problem has been considered by Box and Hill (1967) and many other authors, as reviewed by Hill (1978) and Rippin (1988). The present analysis is relevant to such studies, as will be indicated, but our emphasis is on model discrimination with existing data.

### Model Discrimination with Existing Data

Several approaches to this problem have appeared in the literature. Tschernitz et al. (1946) fitted 18 mechanistic models to their reactor data, eliminated those whose parameter estimates were incompatible with chemical theory, and chose the better-fitting of the two remaining models. Lumpkin et al. (1969) tested a larger candidate set and reported that several additional models fitted well; they also showed that the discrimination would be enhanced by including nonisothermal experiments.

Two statistical approaches to model discrimination are prevalent, as discussed by Chow (1981): (i) seeking the best predictor according to the data, and (ii) seeking the most probable model according to the data. In the first category are methods using the \( C_p \) statistic (Mallows, 1964, 1973; Gorman and Toman, 1966; Daniel and Wood, 1980) for models linear in the parameters, or the information criterion of Akaike (1974) based on divergences from a comprehensive reference model. In the second category are Bayesian approaches, preferably assisted by criticism via sampling theory as advocated by Box (1980). The latter path is taken here as a natural way to seek a good mechanistic model, with the attendant benefits of better understanding and a physicochemical basis for any extrapolations needed.

Consider a set of observations \( y_1, \ldots, y_n \) of a single response variable, obtained or to be obtained in independent experiments \( u = 1, \ldots, n \) on a chemical process. Let \( \xi_u \) denote the setting of the vector \( \xi \) of independent variables (temperature, pressure, initial concentrations, etc.) in experiment \( u \). Then the response \( y_u \) consists of an expectation value \( E(y_u | \xi_u) \) plus an error \( e_u \). Inserting a model \( f_j(\xi, \theta) \) for the expectation values, we get the representation

\[
y_u = f_j(\xi_u, \theta) + e_u \quad u = 1, \ldots, n
\]

for existing or future observations. For existing observations \( y_u \), this equation defines the errors \( e_u \) as functions of the parameter vector \( \theta \), on the postulate that the expectation model form \( f_j \) is true. For future observations, the distribution of possible values of each \( y_u \) will be modeled by the same equation with \( \theta \) fixed, and with \( e_u \) a random variable simulated by sampling from a Normal error distribution \( N(0, \sigma^2_u) \). Each model \( f_j \) is assumed to be differentiable with respect to its parameters.

The distribution of error becomes simpler if we normalize Eqs. 2 to a uniform precision. Let \( w_u = \sigma^2_u/\sigma^2 \) be the specified ratio of the precision of observations of \( y \) at \( \xi_u \) to a standard precision \( 1/\sigma^2 \); then multiplication of Eqs. 2 by \( w_u^{1/2} \) gives the expressions

\[
Y_u = \mathcal{F}(\xi_u, \theta) + E_u \quad u = 1, \ldots, n
\]

for the weighted observations \( Y_u = y_u \sqrt{w_u} \) in terms of the weighted expectation functions \( \mathcal{F}(\xi, \theta) = f_j(\xi, \theta) \sqrt{w_u} \) and weighted errors \( E_u = e_u \sqrt{w_u} \). Then model \( E_1, \ldots, E_n \) as independent samples from the distribution \( N(0, \sigma^2) \), whether analyzing data or simulating future observations.

Consider a list \( (M_1, \ldots, M_j) \) of candidate expectation models. We assign probabilities \( p(M_1), \ldots, p(M_j) \) to these models \( a \ priori \), that is to say, without any reference to the observations. It would often be reasonable to choose these values to be equal. These probabilities need not add up to 1, since only their relative values affect the outcome.

If the models were completely specified, their ranking on the data could be obtained by straightforward application of Bayes' theorem (Bayes, 1763; Box and Tiao, 1975). However, the models considered here contain unknown parameters, and consequently we need a prior probability density for the parameter vector \( \theta \) of each model \( M_j \). An impartial method of choosing such priors was proposed by Box and Henson (1969, 1970); an improved development of it is included here.

### Case I: Variance known

For a given experimental design, let \( p(Y | M_j, \sigma) \) denote the probability density of weighted observation vectors \( Y = (Y_1, \ldots, Y_n)^T \) predicted by Eq. 3 with each \( E_u \) distributed independently as \( N(0, \sigma^2) \). Then according to Bayes' theorem, the posterior probability of model \( M_j \) conditional on the actual data \( Y \) and a given variance \( \sigma^2 \) is

\[
p(M_j | Y, \sigma) = p(M_j) p(Y | M_j, \sigma) / C,
\]

in which \( C \) is a normalization constant, equal for all the models.
If the \( j \)th model contains an unknown parameter vector \( \theta_j \), then integration over all values of this vector gives the following form of Eq. 4,

\[
p(M_j | Y, \sigma) \propto p(M_j) \int p(Y | \theta_j, M_j, \sigma) p(\theta_j | M_j) d\theta_j
\]  

(5)

with the same proportionality constant for every model. If \( p(\theta | M_j) \) is nearly uniform over the region of \( \theta_j \) in which \( p(Y | \theta_j, M_j, \sigma) \) is appreciable for the given \( Y \), then Eq. 5 reduces to

\[
p(M_j | Y, \sigma) \propto p(M_j) \int p(Y | \theta_j, M_j, \sigma) d\theta_j
\]  

(6)

The last integrand takes the form

\[
p(Y | \theta_j, M_j, \sigma) = \prod_{u=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} \left[ Y_u - \mathcal{F}_j(\xi_u, \theta_j) \right]^2 \right\}
\]

at each value of \( \theta_j \) when the error term in each of Eqs. 3 is treated as an independent sample from the Normal distribution \( N(0, \sigma^2) \). In the sample space of data and residuals thus generated, \( S_j/\sigma^2 \) is a random variable distributed as \( \chi^2 \) with \( n - p_j \) degrees of freedom. Over this space, the exponential term in Eq. 9 has the expectation

\[
E[\exp(-\tilde{S}_j/2\sigma^2)] = \int_0^\infty \exp(-x^2/2) p(\chi^2 | n - p_j) \, dx = \frac{\Gamma((n-p)/2)}{\Gamma((n-1)/2)} \]  

(10)

and \( p(M_j | Y, \sigma) \) has expectation \( p(M_j) \). The formula

\[
p(\theta_j | M_j) \propto \tilde{X}_j^T \tilde{X}_j \left[ \tilde{X}_j^T \tilde{X}_j + \frac{1}{\sigma^2} \right]^{-1/2}(2\pi\sigma^2)^{-(n - p_j)/2} \exp\left( -\tilde{S}_j/2\sigma^2 \right)
\]

(9)

when the region of integration in Eq. 8 is treated as unbounded. The values \( |\tilde{X}_j^T \tilde{X}_j|, p_j, \tilde{S}_j, \tilde{\theta}_j \) are readily computable with modern software.

Now consider the predictive distribution of the members of Eq. 9 over conceptual replications of the experimental program and model fitting. Postulate the model \( M_j \) fitted to the existing data to be true, and the errors in the conceptual observations to be random samples from the Normal distribution \( N(0, \sigma^2) \). The formula

\[
p(0, | M_j) = \text{Const. for all } j
\]

for \( p(0, | M_j) \) then follows when one takes the expectation of each member of Eq. 9 over the foregoing sample space. The quantity \( p(\theta_j | M_j) \) thus obtained is a prior density relative to the conceptual sampling process just described. Box and Henson (1969, 1970) gave a similar formula for \( p(0, | M_j) \), but viewed it as an expectation before the taking of any data. The present view is required for consistency of Eq. 11 with the least-squares analysis of the actual data.

Equations 9 and 11 gives the posterior probabilities of the candidate models as

\[
\int p(Y | \theta_j, M_j, \sigma) \, d\theta_j = \int x^2 | \tilde{X}_j^T \tilde{X}_j \left[ \tilde{X}_j^T \tilde{X}_j + \frac{1}{\sigma^2} \right]^{-1/2}(2\pi\sigma^2)^{-(n - p_j)/2} \exp\left( -\tilde{S}_j/2\sigma^2 \right)
\]

(8)

and the posterior probability of model \( M_j \) becomes

\[
p(M_j | Y, \sigma) \propto p(M_j) p(\theta_j | M_j) \]

\[
\left| \tilde{X}_j^T \tilde{X}_j \right|^{-1/2}(2\pi\sigma^2)^{-(n - p_j)/2} \exp\left( -\tilde{S}_j/2\sigma^2 \right)
\]

(9)

after removal of the common factor \( 2^{n/2} \). Finally, an optional normalization over the candidates gives

\[
\pi(M_j | Y, \sigma) = p(M_j | Y, \sigma) / \sum_k p(M_k | Y, \sigma)
\]

(13)

The index \( r \) is numbered here over those parameters of model \( M_j \) that are estimated by unconstrained least-squares conditions \( \partial S_j / \partial \theta_j = 0 \). Any indeterminate parameters, and any determined by constraints (such as nonnegativity in Example 3), are not counted in Eq. 7b but are included in the vector \( \tilde{\theta}_j \) at their last values reached in the least-squares computation. With this convention, \( \tilde{X}_j^T \tilde{X}_j \) is a positive definite matrix of order \( p_j \). The integral in Eq. 6 (taken over the space \( \theta^\sigma \) with coordinates \( \theta_{j1}, \ldots, \theta_{jp_j} \)) then can be approximated as

\[
\int p(Y | \theta_j, M_j, \sigma) \, d\theta_j
\]
as the posterior probability share held by model \( M_j \) according to the data used.

The factor \( 2^{-p_i/2} \) in Eq. 12 may be viewed as a penalty factor for the number of parameters in a model. It offsets the improvement in \( \exp(-\frac{S_j^2}{2\sigma^2}) \) to be expected if a parameter-free model were augmented with \( p_i \) worthless parameters, each of which merely removed one residual degree of freedom. Omission of this factor would replace Eq. 12 by a likelihood-ratio selection criterion, which is known to favor overparameterized models (Reilly, 1970; Chow, 1981). Use of Eq. 12 avoids this difficulty.

**Case II: Variance unknown, but estimate available from data**

Suppose that \( \sigma^2 \) is unknown, but that there is genuine replication of at least some of the runs. Let these replications supply a variance estimate \( s^2 = S_e/\nu_e \) having \( \nu_e \) degrees of freedom. Then the residual sums of squares \( \hat{S}_j \) for the models take the forms

\[
\hat{S}_1 = S_1^j + S_e \\
\hat{S}_2 = S_2^j + S_e \\
\vdots \\
\hat{S}_j = S_j^j + S_e,
\]

where \( S_1^j, S_2^j, \ldots, S_j^j \) are the “lack-of-fit” sums of squares remaining after the same “pure error” contribution \( S_e \) is subtracted from each residual sum of squares \( \hat{S}_j \).

In this case, we can take out the common factor \( \exp(-\frac{S_j^2}{2\sigma^2}) \) from each posterior probability in Eq. 12 and obtain

\[
p(M_j | Y, \sigma) \propto p(M_j)2^{-p_i/2} \exp\left(-\frac{S_j^2}{2\sigma^2}\right). \tag{14}
\]

The unknown parameter \( \sigma \) can be removed by integration,

\[
p(M_j | Y, S_e, \nu_e) = \int_0^\infty p(M_j | Y, \sigma) p(\sigma | S_e, \nu_e) d\sigma \tag{15}
\]

with the conditional density function (Box and Tiao, 1973, p. 100)

\[
p(\sigma | S_e, \nu_e) = \frac{2}{\sigma^{\nu_e/2} \Gamma(\nu_e/2)} \exp\left(-\frac{S_e}{2\sigma^2}\right). \tag{16}
\]

Equations 14, 15, and 16 give the posterior probabilities of the candidate models as

\[
p(M_j | Y, S_e, \nu_e) \propto p(M_j)2^{-p_i/2} \int_0^{\infty} \frac{1}{\sigma^{\nu_e+1}} \exp\left[-(S_e + S_j^2)/2\sigma^2\right] d\sigma \\
\times \alpha p(M_j)2^{-p_e/2} (S_e + S_j^2)^{-\nu_e/2} \\
\alpha p(M_j)2^{-p_e/2} S_j^2/\nu_e \\
\alpha \left[1 + \frac{(n - p_j - \nu_e)F_j}{\nu_e}\right]^{-\nu_e/2} \tag{17}
\]

For large \( \nu_e \), Eq. 17 gives a result like Eq. 12, but with the sample estimate \( s_e^2 \) replacing \( \sigma^2 \). The larger the value of \( \nu_e \), the stronger the discrimination.

**Goodness of Fit**

If the postulated models all fitted the data badly, then the preceding results could be misleading. One could be led to choose a poor model because it was not as bad as the other models. Fortunately, the means for resolving this difficulty are immediately available in the form of goodness-of-fit tests, closely related to the calculations already described.

When \( \sigma^2 \) is known, the goodness of fit of the \( j \)th model can be tested by referring the sample value \( \chi^2 = S_j^2/\sigma^2 \) to the distribution function \( P(\chi^2 | \nu) = 1 - P(\chi^2 | \nu) \), tabulated in Abramowitz and Stegun (1972), yields the probability of obtaining a \( \chi^2 \) value larger than \( S_j^2/\sigma^2 \) in a random replication of the observations and computations, on the hypothesis that model \( M_j \) and the assumed Normal error distribution are true for the experimental situation. Note that the sample value \( S_j^2/\sigma^2 \) also appears in Eq. 12.

When \( \sigma^2 \) is unknown, but a sample estimate \( s^2 \) is available, the goodness of fit of the \( j \)th model can be tested by use of the entries in the following analysis of variance (ANOVA) table:

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>Sum of Squares</th>
<th>Deg of Freedom</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of Fit</td>
<td>( S_j^j )</td>
<td>( n - p_j - \nu_e )</td>
<td>( S_j^j/(n - p_j - \nu_e) = s_j^2 )</td>
</tr>
<tr>
<td>Pure Error</td>
<td>( S_e )</td>
<td>( \nu_e )</td>
<td>( s_e^2 )</td>
</tr>
</tbody>
</table>

A test of the hypothesis of adequacy of the \( j \)th fitted model can be made by referring the ratio \( F_j = s_j^2/s^2 \) to a table of critical values \( F(\nu_1, \nu_2; Q) \) at significance level \( Q \) with \( \nu_1 = (n - p_j - \nu) \) and \( \nu_2 = \nu_2 \) degrees of freedom. Fuller information can be obtained by evaluating the probability \( Q(F_j | \nu_1, \nu_2) \) by interpolation in the same table, or via a Fortran function as in Press et al. (1992). The value \( Q(F_j | \nu_1, \nu_2) \) is the probability of obtaining an \( F \) value larger than \( F_j \) in random sampling, on the hypothesis that model \( M_j \) and the assumed Normal error distribution are true for the experimental situation. This test was originally derived from sampling theory, but follows equally well from a Bayesian analysis (Box and Tiao, 1965).

Insertion of \( F_j = s_j^2/s^2 \) into Eq. 17 gives

\[
p(M_j | Y, S_e, \nu_e) \propto p(M_j)2^{-p_i/2} \left[1 + \frac{(n - p_j - \nu_e)F_j}{\nu_e}\right]^{-\nu_e/2}. \tag{18}
\]

Thus, the variance ratio \( F_j \) appears in the posterior probability for model \( M_j \) as well as in the ANOVA test.

If the variance of \( y \) is unknown and no replicate observations are available, one may seek an estimate from mass-balance residuals as in Stewart and Mastenbrook (1984), or from the residuals found by fitting a high-order model as in Example 3 below.
Sequential Analysis

After computing the posterior probabilities for the initial set of \( n \) observations, one may want to strengthen the discrimination by taking additional data. Our analysis holds directly for such extended data sets, as if the experiments were all included in the original plan.

Equation 12 differs from the posterior density expression of Box and Hill (1967) because our prior densities \( p(\theta_1 | M_1) \) are calculated with Eq. 11, using the data available at the time. Equation 12 is appropriate when \( \sigma^2 \) is given, whereas Eq. 17 must be used when the variance information comes from experiments.

Examples

Three examples are provided here to illustrate the use of the formulas. Example 1 treats linear models for data with a given variance; Example 2 treats a pair of nonlinear models, using a variance estimated by replication; and Example 3 treats a set of eighteen nonlinear models, using a variance estimated from residuals of a higher-order model.

Example 1

Consider two models, one included in the other, that are to be tested with data of known variance \( \sigma^2 \) from an orthogonal experimental design in the independent variables \( \xi_1 \) and \( \xi_2 \):

\[
\mathbb{F}_1(\xi, \theta_1) = \theta_1 \xi_1; \quad \mathbb{F}_2(\xi, \theta_2) = \theta_1 \xi_1 + \theta_2 \xi_2.
\]

The orthogonality relation \( \sum_{i=1}^{n} \xi_{1i} \xi_{2i} = 0 \) of the experimental design makes the least-squares estimates of \( \theta_1 \) identical for the two models; therefore, the notation \( \theta_{\mathbb{F}} \) of Eq. 7b is abbreviated here to \( \theta \). The orthogonality also yields the relation \( \langle \hat{S}_1 - \hat{S}_2 \rangle = \theta_2^2 \sum_{i=1}^{n} \xi_{2i}^2 \) between the residual sums of squares for these two models. With \( p(M_1) = p(M_2) \), Eq. 12 gives

\[
\frac{p(M_1 | Y, \sigma)}{p(M_2 | Y, \sigma)} = 2^{1/2} \exp \left( \frac{\langle \hat{S}_2 - \hat{S}_1 \rangle}{2 \sigma^2} \right)
\]

\[
= 2^{1/2} \exp \left[ -\frac{\theta_2^2}{2} \sum_{i=1}^{n} \xi_{2i}^2 / 2 \sigma^2 \right]
\]
as the ratio of posterior probabilities. This ratio never exceeds \( \sqrt{2} \).

If \( \theta_2 \) is nonzero, then only Model 2 is true. The last expression for the probability ratio will favor Model 2 as soon as enough data are taken to make the exponential factor smaller than \( 1/\sqrt{2} \). The smaller the true value of \( \theta_2 \), the greater the number of experiments required to establish the appropriateness of Model 2.

If Model 1 is true, then both models are true and \( \theta_2 \) is zero. The estimate \( \hat{\theta}_2 \) then has expectation zero, and \( \langle \hat{S}_1 - \hat{S}_2 \rangle / \sigma^2 \) is distributed as \( \chi^2 \) with 1 degree of freedom. The resulting probability ratio has expectation \( 2^{1/2} \mathbb{E}[\exp(\chi^2/2)] \) if \( \sigma^2 \) is known, computed by the method in Eq. 10. In this case, the probability ratio to be obtained from a future set of experiments is equally likely to favor either model. The lack of a trend in the ratio with increasing \( n \) would indicate the choice to be a hair-splitting one, and the simpler model would be preferred on grounds of parsimony.

In general, if the most probable model includes another with comparable posterior probability and adequate goodness of fit, the choice of the simpler model should be considered. A reasonable procedure would be to choose whichever model wins in the majority of comparisons, when the posterior probability ratio is calculated for various data sets obtained from further experiments or from random resampling of the data at hand. Ties or close contests would be resolved in favor of the simpler model.

Example 2

Suppose that some data are available for the concentration \( [B] \) of species B as a function of time in a constant-volume, isothermal batch reactor containing chemical species A, B, and C. Sixteen simulated pairs of replicate observations similar to those considered by Box and Coutie (1956) are given in Table 1 for the initial condition \([A]_0 = 1, [B]_0 = [C]_0 = 0\). The following two models (derived by integration of the corresponding differential equations) are postulated.

\begin{align*}
\text{Model 1. Consecutive first-order reactions} \quad & k_1 B \rightarrow C, \\
\text{giving} \quad & [B] = \frac{k_1}{k_2 - k_1} \left( \exp(-k_1 t) - \exp(-k_2 t) \right).
\end{align*}

\begin{align*}
\text{Model 2. Parallel first-order reactions} \quad & k_1' A \rightarrow B \rightarrow C, \\
\text{and} \quad & k_2' A \rightarrow C, \\
\text{giving} \quad & [B] = \frac{k_1'}{\lambda_2 - \lambda_3} \left( \exp(-\lambda_2 t) - \exp(-\lambda_3 t) \right) \\
& \lambda_2 = (p + q)/2, \\
& \lambda_3 = (p - q)/2, \\
& p = k_1' + k_2' + k_3', \\
& q = (p^2 - 4k_2'k_3')^{1/2}.
\end{align*}

Fitting these models to the data by nonlinear least squares yields the results in Table 2. Model 1 fits the data better than Model 2 and has a much higher posterior probability. The posterior probabilities here are calculated from Eq. 17 with \( p(M_1) = p(M_2) \) and normalization to a posterior sum of 1.

<table>
<thead>
<tr>
<th>Table 1. Simulated Data for Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t, \text{ min} )</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>120</td>
</tr>
<tr>
<td>140</td>
</tr>
</tbody>
</table>
Analyses of variance for both models are given in Table 3 and show a significant lack of fit for Model 2. These results are as expected, since the data were constructed from Model 1 augmented with simulated random errors.

### Example 3

Tschernitz et al. (1946) reported a detailed study of the kinetics of hydrogenation of mixed isooctenes over a supported nickel catalyst. Their article gives 40 unreplicated observations of the hydrogenation rate. The independent variables investigated were the reaction temperature $T$ and the partial pressures $p_H$, $p_U$, and $p_S$ of hydrogen, unsaturates (mixed isooctenes), and saturated products (isooctanes). Eighteen rival models of the reaction mechanism were formulated and fitted by least squares to the experimental data. The data and the models are reanalyzed here.

For the present study, the observations were expressed as values of the response function $y = \ln \vartheta$. Here $\vartheta$ is the hydrogenation rate in mols per hour per unit mass of catalyst, adjusted to a standard level of catalyst activity. Weights were assigned to these observations according to the formula

$$\sigma_j^2 = (0.00001/\Delta n_{RI})^2 + (0.1)^2$$

based on assigned standard deviations of 0.00001 for the observations of refractive-index difference $\Delta n_{RI}$ (used in calculating the reactant conversions) and 0.1 for the catalyst activity corrections. The activity variance, $(0.1)^2$, proved to be dominant in Eq. 19 over the range of the experiments; thus the weights $w_u = 1/(\sigma_j^2)$ were nearly equal for the 40 experiments.

The eighteen models investigated by Tschernitz et al. are expressed in Table 4 as expectation models for their desk calculations, but used weights of 1, which are inappropriate for the resulting transformed response functions. For comparison, we have refitted the data with the same linearized models, but with different weights $w_u$ derived for each

<table>
<thead>
<tr>
<th>Table 2. Model Fitting Results for Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

### Table 3. Analyses of Variance for Example 2

<table>
<thead>
<tr>
<th>ANOVA for Model 1: Source of Variance</th>
<th>Sum of Squares</th>
<th>Deg. of Freedom</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>0.114243</td>
<td>34</td>
<td>0.004396</td>
</tr>
<tr>
<td>Lack of fit</td>
<td>0.070335</td>
<td>16</td>
<td>0.002439</td>
</tr>
<tr>
<td>Pure error</td>
<td>0.043908</td>
<td>18</td>
<td>0.002439</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA for Model 2: Source of Variance</th>
<th>Sum of Squares</th>
<th>Deg. of Freedom</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>0.174806</td>
<td>33</td>
<td>0.008727</td>
</tr>
<tr>
<td>Lack of fit</td>
<td>0.130898</td>
<td>15</td>
<td>0.002439</td>
</tr>
<tr>
<td>Pure error</td>
<td>0.043908</td>
<td>18</td>
<td>0.002439</td>
</tr>
</tbody>
</table>
Table 4. Expectation Models for Response $y = \ln \theta$ in Example 3

| Model | Residual Sum of Squares | Posterior Probability $P(M_i | Y, S_e, \nu_e)$ | Lack-of-Fit $L_i$ | Exp. Error $E_i$ |
|-------|-------------------------|---------------------------------------------|-----------------|-----------------|
| 1     | 970.2                   | 0.000                                       | 13.4            | 19              |
| 2     | 2,156.7                 | 0.000                                       | 29.3            | 27              |
| 3     | 279.5                   | 0.014                                       | 3.6             | 13              |
| 4     | 192.2                   | 0.167                                       | 2.4             | 17              |
| 5     | 1,013.8                 | 0.000                                       | 14.0            | 19              |
| 6     | 2,586.3                 | 0.000                                       | 33.6            | 21              |
| 7     | 211.1                   | 0.213                                       | 2.3             | 18              |
| 8     | 165.2                   | 0.605                                       | 1.9             | 15              |
| 9     | 970.2                   | 0.000                                       | 13.4            | 20              |
| 10    | 844.9                   | 0.000                                       | 11.5            | 19              |
| 11    | 838.2                   | 0.000                                       | 11.9            | 18              |
| 12    | 1,013.8                 | 0.000                                       | 14.0            | 19              |
| 13    | 767.5                   | 0.000                                       | 10.4            | 19              |
| 14    | 788.8                   | 0.000                                       | 11.3            | 18              |
| 15    | 420.1                   | 0.000                                       | 5.9             | 17              |
| 16    | 485.1                   | 0.000                                       | 6.2             | 19              |
| 17    | 2,156.7                 | 0.000                                       | 29.3            | 23              |
| 18    | 925.4                   | 0.000                                       | 11.5            | 21              |

Table 5. Testing of Kinetic Models*

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>Residual</th>
<th>Posterior</th>
<th>Lack-of-Fit</th>
<th>Exp. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>970.2</td>
<td>0.000</td>
<td>13.4</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>2,156.7</td>
<td>0.000</td>
<td>29.3</td>
<td>27</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>279.5</td>
<td>0.014</td>
<td>3.6</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>192.2</td>
<td>0.167</td>
<td>2.4</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>1,013.8</td>
<td>0.000</td>
<td>14.0</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>2,586.3</td>
<td>0.000</td>
<td>33.6</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>211.1</td>
<td>0.213</td>
<td>2.3</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>165.2</td>
<td>0.605</td>
<td>1.9</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>970.2</td>
<td>0.000</td>
<td>13.4</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>10</td>
<td>844.9</td>
<td>0.000</td>
<td>11.5</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>11</td>
<td>838.2</td>
<td>0.000</td>
<td>11.9</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>12</td>
<td>1,013.8</td>
<td>0.000</td>
<td>14.0</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>13</td>
<td>767.5</td>
<td>0.000</td>
<td>10.4</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>14</td>
<td>788.8</td>
<td>0.000</td>
<td>11.3</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>15</td>
<td>420.1</td>
<td>0.000</td>
<td>5.9</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>16</td>
<td>485.1</td>
<td>0.000</td>
<td>6.2</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>17</td>
<td>2,156.7</td>
<td>0.000</td>
<td>29.3</td>
<td>23</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>925.4</td>
<td>0.000</td>
<td>11.5</td>
<td>21</td>
<td>17</td>
</tr>
</tbody>
</table>

*For data of Tschernitz et al. (1946).

Adapted from Box and Henson (1969, 1970), but with $p(\theta_i | M_i)$ regarded as an expectation before the taking of any data, a view contradicted by Eq. 11 and questioned by Kneermas (1973). This difficulty is resolved here by evaluating $p(\theta_i | M_i)$ after the data are fitted, as an expectation over a sample space of replicate observations generated by Eq. 3 with $\theta_i = \theta_j$, and errors distributed as $N(0, \sigma^2)$. By this method, we find that Eqs. 12 and 17 hold directly for sequentially investigating of rival models. The first author introduced these changes.

Bayes' theorem and sampling theory were both essential in this work. Both were used in deriving Eq. 11 and the subsequent posterior density formulas for choosing the most probable model; either approach yields the $\chi^2$ and $F$ criteria for testing the adequacy of each model. This outcome is consistent with the conclusion of Box (1980), that Bayesian inference and sampling theory are both essential in modeling investigations.

The factor $2^{-p/2}$ in Eqs. 12 and 17 comes from the calculation of $p(\theta_i | M_i)$ as shown in Eq. 11. This factor facilitates the selection of a parsimoniously parameterized model, thus correcting a reported weakness of previous Bayesian discrimination methods (Reilly, 1970; Chow, 1981).

Example 3 illustrates several practical points:

1. By use of a constrained least-squares algorithm, physically acceptable parameter estimates were obtained for every candidate model, leaving the discrimination to be done by means of posterior probabilities and ANOVA tests.
2. Weighting of the observations is inherent in any least-squares procedure, and needs to be based on a list or model of the expected relative precisions of the observations.
3. Rearrangements of kinetic models into linear forms, though useful for graphical analysis, are unnecessary for parameter estimation with modern algorithms. Such rearrangements make discrimination more difficult, by requiring model-dependent transformations of the data, the weights, and the residuals to get valid comparisons of models.
4. Replicate experiments are very important for estimation of the experimental error statistics, $S_e$ and $\nu_e$. Residuals of rigorous constraints, such as mass balances, are also useful for this purpose. When such information is lacking, approximations of $S_e$ and $\nu_e$ may be obtainable by high-order polynomial regression of the observations, or of the residuals for the best-fitting model.

Acknowledgment

The authors express their gratitude for financial support received from the Office of Naval Research, under Contract Nonr-1202(17), Project Number 042-222, and from the National Science Foundation, under Grants GK-1055 and CTS-9120325, while carrying out this work.

Notation

$p(M_i)$ = prior probability assigned to model $M_i$
$p(x | y) =$ density of $x^2$ distribution with $v$ degrees of freedom
$S_e = S_e/(n-p_1-\nu_e)$, lack-of-fit variance for model $M_j$
$\nu =$ observation number
$X = $ matrix of parametric sensitivities for model $M_i$; see Eq. 7b
$\hat{X}^T \hat{X} = $ determinant of matrix $\hat{X}^T \hat{X}$
$Y = $ vector of weighted observations
$\Gamma(x) = $ gamma function at $x$
$\sigma^2 = $ variance of observations at $\xi_i$
$\Sigma^2 = $ transpose of vector or matrix
$\sigma^2 = $ least-squares value
$= $ defines the preceding symbol by the next expression
Literature Cited


Manuscript received Jan. 8, 1996, and revision received May 2, 1996.