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 $\boldsymbol{\xi}$ of the experimental conditions is given exactly by

$$E(y) = f(\boldsymbol{\xi}, \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta} := (\theta_1, \theta_2, \ldots, \theta_p)^T$ 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 $f(\boldsymbol{\xi}, \boldsymbol{\theta})$ 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 $\boldsymbol{\xi}$ it would often be true that the relationship between $E(y)$ and $\boldsymbol{\xi}$ was smooth and could be locally *approximated* by an interpolation function, $g(\boldsymbol{\xi}, \boldsymbol{\theta})$. Then $g(\boldsymbol{\xi}, \boldsymbol{\theta})$ 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-

Present address of T. L. Henson: Osram Sylvania Inc., Towanda, PA 18848.

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 ξ_i. 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 could lead to important developments.

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 Coutie, 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 $\boldsymbol{\xi}_u$ denote the setting of the vector $\boldsymbol{\xi}$ of independent variables (temperature, pressure, initial concentrations, etc.) in experiment u. Then the response y_u consists of an expectation value $E(y \mid \boldsymbol{\xi}_u)$ plus an error ε_u. Inserting a model $f_j(\boldsymbol{\xi}, \boldsymbol{\theta}_j)$ for the expectation values, we get the representation

$$y_u = f_j(\boldsymbol{\xi}_u, \boldsymbol{\theta}_j) + \varepsilon_u \qquad u = 1, \ldots, n \tag{2}$$

for existing or future observations. For existing observations y_u, this equation defines the errors ε_u as functions of the parameter vector $\boldsymbol{\theta}_j$, 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 $\boldsymbol{\theta}_j$ fixed, and with ε_u a random variable simulated by sampling from a Normal error distribution $N(0, \sigma_u^2)$. 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 $w_u = \sigma^2/\sigma_u^2$ be the specified ratio of the precision of observations of y at $\boldsymbol{\xi}_u$ to a standard precision $1/\sigma^2$; then multiplication of Eqs. 2 by $\sqrt{w_u}$ gives the expressions

$$Y_u = \mathfrak{F}_j(\boldsymbol{\xi}_u, \boldsymbol{\theta}_j) + \mathcal{E}_u \qquad u = 1, \ldots, n \tag{3}$$

for the weighted observations $Y_u := y_u\sqrt{w_u}$ in terms of the weighted expectation functions $\mathfrak{F}_j(\boldsymbol{\xi}_u, \boldsymbol{\theta}_j) := f_j(\boldsymbol{\xi}_u, \boldsymbol{\theta}_j)\sqrt{w_u}$ and weighted errors $\mathcal{E}_u := \varepsilon_u\sqrt{w_u}$. Then we model $\mathcal{E}_1, \ldots, \mathcal{E}_n$ 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, 1973). However, the models considered here contain unknown parameters, and consequently we need a prior probability density for the parameter vector $\boldsymbol{\theta}_j$ 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 $p(\boldsymbol{Y} \mid M_j, \sigma)$ denote the probability density of weighted observation vectors $\boldsymbol{Y} := (Y_1, \ldots, Y_n)^T$ predicted by Eq. 3 with each $\mathcal{E}_u$ distributed independently as $N(0, \sigma^2)$. Then according to Bayes' theorem, the posterior probability of model M_j conditional on the actual data $\boldsymbol{Y}$ and a given variance σ^2 is

$$p(M_j \mid \boldsymbol{Y}, \sigma) = p(M_j)p(\boldsymbol{Y} \mid M_j, \sigma)/C, \tag{4}$$

in which C is a normalization constant, equal for all the models.

If the jth model contains an unknown parameter vector $\boldsymbol{\theta}_j$, then integration over all values of this vector gives the following form of Eq. 4,

$$p(M_j \mid \boldsymbol{Y}, \sigma) \propto p(M_j) \int p(\boldsymbol{Y} \mid \boldsymbol{\theta}_j, M_j, \sigma) p(\boldsymbol{\theta}_j \mid M_j) d\boldsymbol{\theta}_j \quad (5)$$

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

$$p(M_j \mid \boldsymbol{Y}, \sigma) \propto p(M_j) p(\boldsymbol{\theta}_j \mid M_j) \int p(\boldsymbol{Y} \mid \boldsymbol{\theta}_j, M_j, \sigma) d\boldsymbol{\theta}_j. \quad (6)$$

The last integrand takes the form

$$p(\boldsymbol{Y} \mid \boldsymbol{\theta}_j, M_j, \sigma) = \prod_{u=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2} [Y_u - \mathfrak{F}_j(\xi_u, \boldsymbol{\theta}_j)]^2 \right\}$$

at each value of $\boldsymbol{\theta}_j$ when the error term in each of Eqs. 3 is treated as an independent sample from the Normal distribution $N(0, \sigma^2)$.

Local linearization of each $\mathfrak{F}_j$ term with respect to the parameters, at the least-squares point $\hat{\boldsymbol{\theta}}_j$ for the given $\boldsymbol{Y}$, gives the approximation

$$p(\boldsymbol{Y} \mid \boldsymbol{\theta}_j, M_j, \sigma) \approx (2\pi\sigma^2)^{-n/2} \times \exp\left\{ -\frac{1}{2\sigma^2} \left[\hat{S}_j + \left(\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j\right)^T \hat{\boldsymbol{X}}_j^T \hat{\boldsymbol{X}}_j \left(\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j\right) \right] \right\}, \quad (7)$$

consistent with the Gaussian normal equations for nonlinear least squares (Gauss, 1809; Bates and Watts, 1988). Here $\hat{S}_j := S(\hat{\boldsymbol{\theta}}_j)$ is the minimum sum of squares (also called the residual sum of squares) for model M_j,

$$\hat{S}_j = \sum_{u=1}^{n} \left[Y_u - \mathfrak{F}_j\left(\xi_u, \hat{\boldsymbol{\theta}}_j\right) \right]^2 \quad (7a)$$

and $\hat{\boldsymbol{X}}_j$ is the $n \times p_j$ matrix with elements

$$\left[\hat{\boldsymbol{X}}_j\right]_{ur} = \frac{\partial \mathfrak{F}_j(\xi_u, \boldsymbol{\theta}_j)}{\partial \theta_{jr}} \text{ evaluated at } \boldsymbol{\theta}_j = \hat{\boldsymbol{\theta}}_j \left\{ \begin{matrix} u = 1, \ldots, n \\ r = 1, \ldots, p_j \end{matrix} \right\}. \quad (7b)$$

The index r is numbered here over those parameters of model M_j that are estimated by unconstrained least-squares conditions $\partial S/\partial \theta_{jr} = 0$. 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 $\hat{\boldsymbol{\theta}}_j$ at their last values reached in the least-squares computation. With this convention, $\hat{\boldsymbol{X}}_j^T \hat{\boldsymbol{X}}_j$ is a positive definite matrix of order p_j. The integral in Eq. 6 (taken over the space $\boldsymbol{\theta}_j^e$ with coordinates $\theta_{j1}, \ldots, \theta_{jp_j}$) then can be approximated as

$$\int p(\boldsymbol{Y} \mid \boldsymbol{\theta}_j, M_j, \sigma) d\boldsymbol{\theta}_j^e \approx |\hat{\boldsymbol{X}}_j^T \hat{\boldsymbol{X}}_j|^{-1/2} (2\pi\sigma^2)^{-(n-p_j)/2} \exp\left(-\hat{S}_j/2\sigma^2\right), \quad (8)$$

and the posterior probability of model M_j becomes

$$p(M_j \mid \boldsymbol{Y}, \sigma) \propto p(M_j) p(\boldsymbol{\theta}_j \mid M_j) |\hat{\boldsymbol{X}}_j^T \hat{\boldsymbol{X}}_j|^{-1/2} (2\pi\sigma^2)^{-(n-p_j)/2} \exp\left(-\hat{S}_j/2\sigma^2\right) \quad (9)$$

when the region of integration in Eq. 8 is treated as unbounded. The values $|\hat{\boldsymbol{X}}_j^T \hat{\boldsymbol{X}}_j|$, p_j, $\hat{S}_j$, and $\hat{\boldsymbol{\theta}}_j$ 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 $N(0, \sigma^2)$. In the sample space of data and residuals thus generated, $\hat{S}_j/\sigma^2$ is a random variable distributed as χ^2 with $n - p_j$ degrees of freedom. Over this space, the exponential term in Eq. 9 has the expectation

$$\begin{aligned} \mathrm{E}\left[\exp\left(-\hat{S}_j/2\sigma^2\right)\right] &= \int_0^\infty \exp(-\chi^2/2) p\left(\chi^2 \mid n - p_j\right) d\chi^2 \\ &= \int_0^\infty \exp(-\chi^2) \frac{(\chi^2)^{(\nu-2)/2}}{\Gamma(\nu/2) 2^{\nu/2}} d\chi^2 \\ &= 2^{-(n-p_j)/2} \quad \text{with} \quad \nu = n - p_j \end{aligned} \quad (10)$$

and $p(M_j \mid \boldsymbol{Y}, \sigma)$ has expectation $p(M_j)$. The formula

$$p(\boldsymbol{\theta}_j \mid M_j) |\hat{\boldsymbol{X}}_j^T \hat{\boldsymbol{X}}_j|^{-1/2} (2\pi\sigma^2)^{-(n-p_j)/2} 2^{-(n-p_j)/2} = \text{Const. for all } j \quad (11)$$

for $p(\boldsymbol{\theta}_j \mid M_j)$ then follows when one takes the expectation of each member of Eq. 9 over the foregoing sample space.

The quantity $p(\boldsymbol{\theta}_j \mid M_j)$ thus obtained is a prior density relative to the conceptual sampling process just described. Box and Henson (1969, 1970) gave a similar formula for $p(\boldsymbol{\theta}_j \mid M_j)$, 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

$$p(M_j \mid \boldsymbol{Y}, \sigma) \propto p(M_j) 2^{-p_j/2} \exp\left(-\hat{S}_j/2\sigma^2\right) \qquad j = 1, \ldots, J \quad (12)$$

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

$$\pi(M_j \mid \boldsymbol{Y}, \sigma) = p(M_j \mid \boldsymbol{Y}, \sigma) / \sum_k p(M_k \mid \boldsymbol{Y}, \sigma) \quad (13)$$

as the posterior probability share held by model M_j according to the data used.

The factor $2^{-p_j/2}$ in Eq. 12 may be viewed as a penalty factor for the number of parameters in a model. It offsets the improvement in $\exp(-\hat{S}/2\sigma^2)$ to be expected if a parameter-free model were augmented with p_j 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 σ^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 ν_e degrees of freedom. Then the residual sums of squares $\hat{S}_j$ for the models take the forms

$$\begin{aligned}\hat{S}_1 &= S'_1 + S_e\\ \hat{S}_2 &= S'_2 + S_e\\ &\vdots\\ \hat{S}_J &= S'_J + S_e,\end{aligned}$$

where $S'_1, S'_2, \ldots, S'_J$ 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 $\exp(-S_e/2\sigma^2)$ from each posterior probability in Eq. 12 and obtain

$$p(M_j \mid Y, \sigma) \propto p(M_j) 2^{-p_j/2} \exp\left(-S'_j/2\sigma^2\right). \quad (14)$$

The unknown parameter σ can be removed by integration,

$$p(M_j \mid Y, S_e, \nu_e) = \int_0^\infty p(M_j \mid Y, \sigma)\, p(\sigma \mid S_e, \nu_e)\, d\sigma \quad (15)$$

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

$$p(\sigma \mid S_e, \nu_e) = \frac{2}{2^{\nu_e/2}\Gamma(\nu_e/2)} \frac{S_e^{\nu_e/2}}{\sigma^{\nu_e+1}} \exp(-S_e/2\sigma^2). \quad (16)$$

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

$$\begin{aligned}p(M_j \mid Y, S_e, \nu_e) &\propto p(M_j) 2^{-p_j/2} \int_0^\infty \frac{1}{\sigma^{\nu_e+1}}\\ &\quad \times \exp\left[-(S_e + S'_j)/2\sigma^2\right] d\sigma\\ &\propto p(M_j) 2^{-p_j/2} (S_e + S'_j)^{-\nu_e/2}\\ &\propto p(M_j) 2^{-p_j/2} \hat{S}_j^{-\nu_e/2} \qquad j = 1, \ldots, J. \quad (17)\end{aligned}$$

For large ν_e Eq. 17 gives a result like Eq. 12, but with the sample estimate s^2 replacing σ^2. The larger the value of ν_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 σ^2 is known, the goodness of fit of the jth model can be tested by referring the sample value $\chi_j^2 := \hat{S}_j/\sigma^2$ to the distribution function $P(\chi^2 \mid \nu)$ with $\nu = n - p_j$ degrees of freedom. This function is widely available in statistical tables and software. The related function $Q(\chi^2 \mid \nu) := 1 - P(\chi^2 \mid \nu)$, tabulated in Abramowitz and Stegun (1972), yields the probability of obtaining a χ^2 value larger than $\hat{S}_j/\sigma^2$ 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 $\hat{S}_j/\sigma^2$ also appears in Eq. 12.

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

Source of Variance	Sum of Squares	Deg of Freedom	Mean Square
Lack of Fit	S'_j	$n - p_j - \nu_e$	$S'_j/(n - p_j - \nu_e) = s_j^2$
Pure Error	S_e	ν_e	$S_e/\nu_e = s^2$

A test of the hypothesis of adequacy of the jth fitted model can be made by referring the ratio $F_j = s_j^2/s^2$ to a table of critical values $F(\nu_1, \nu_2 \mid Q)$ at significance level Q with $\nu_1 = (n - p_j - n_e)$ and $\nu_2 = \nu_e$ degrees of freedom. Fuller information can be obtained by evaluating the probability $Q(F_j \mid \nu_1, \nu_2)$ by interpolation in the same table, or via a Fortran function as in Press et al. (1992). The value $Q(F_j \mid \nu_1, \nu_2)$ 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

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

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 $p(\boldsymbol{\theta}_j \mid M_j)$ are calculated with Eq. 11, using the data available at the time. Equation 12 is appropriate when σ^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 σ^2 from an orthogonal experimental design in the independent variables ξ_1 and ξ_2:

$$\mathfrak{F}_1(\boldsymbol{\xi}, \boldsymbol{\theta}_1) = \theta_1 \xi_1; \qquad \mathfrak{F}_2(\boldsymbol{\xi}, \boldsymbol{\theta}_2) = \theta_1 \xi_1 + \theta_2 \xi_2.$$

The orthogonality relation $\Sigma_{u=1}^{n} \xi_{u1} \xi_{u2} = 0$ of the experimental design makes the least-squares estimates of θ_1 identical for the two models; therefore, the notation θ_{jr} of Eq. 7b is abbreviated here to θ_r. The orthogonality also yields the relation $(\hat{S}_1 - \hat{S}_2) = \hat{\theta}_2^2 \Sigma_u \xi_{u2}^2$ between the residual sums of squares for these two models. With $p(M_1) = p(M_2)$, Eq. 12 gives

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

$$= 2^{1/2} \exp\left[-\hat{\theta}_2^2 \sum_u \xi_{u2}^2 / 2\sigma^2\right]$$

as the ratio of posterior probabilities. This ratio never exceeds $\sqrt{2}$.

If θ_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 θ_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 θ_2 is zero. The estimate $\hat{\theta}_2$ then has expectation zero, and $(\hat{S}_1 - \hat{S}_2)/\sigma^2$ is distributed as χ^2 with 1 degree of freedom. The resulting probability ratio has expectation $2^{1/2}\mathrm{E}[\exp(-\chi^2/2) \mid_{\nu=1}] = 1$, 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.

Model 1. Consecutive first-order reactions $A \overset{k_1}{\rightarrow} B \overset{k_2}{\rightarrow} C$, giving

$$[B] = \frac{k_1}{k_2 - k_1}\{\exp(-k_1 t) - \exp(-k_2 t)\}.$$

Model 2. Parallel first-order reactions $A \underset{k_2'}{\overset{k_1'}{\rightleftharpoons}} B$ and $A \overset{k_3'}{\rightarrow} C$, giving

$$[B] = \frac{k_1'}{\lambda_2 - \lambda_3}\{\exp(-\lambda_3 t) - \exp(-\lambda_2 t)\}$$

$$\lambda_2 = (p + q)/2$$

$$\lambda_3 = (p - q)/2$$

$$p = k_1' + k_2' + k_3'$$

$$q = (p^2 - 4k_2'k_3')^{1/2}.$$

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 1. Simulated Data for Example 2

t, min	[B]	t, min	[B]
10	0.192, 0.140	160	0.407, 0.464
20	0.144, 0.240	180	0.439, 0.380
30	0.211, 0.161	200	0.387, 0.393
40	0.423, 0.308	220	0.362, 0.324
60	0.406, 0.486	240	0.269, 0.293
80	0.421, 0.405	260	0.240, 0.424
100	0.457, 0.519	280	0.269, 0.213
120	0.505, 0.537	300	0.297, 0.303
140	0.558, 0.581	320	0.271, 0.223

Table 2. Model Fitting Results for Example 2

Model j	$\hat{k}_{j1}$	$\hat{k}_{j2}$	$\hat{k}_{j3}$	$\hat{S}_j$	$\pi(M_j \mid Y, S_e, \nu_e)$
1	1.21×10^{-2}	6.44×10^{-3}		0.1142	0.987
2	1.58×10^{-2}	7.8×10^{-3}	7.8×10^{-3}	0.1748	0.013

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 $y := \ln \mathcal{R}$. Here $\mathcal{R}$ 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 adjusted observations according to the formula

$$\sigma_y^2 = (0.00001/\Delta n_{\mathrm{RI}})^2 + (0.1)^2 \tag{19}$$

based on assigned standard deviations of 0.00001 for the observations of refractive-index difference Δn_{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 $w_u = 1/(\sigma_y^2)_u$ were nearly equal for the 40 experiments.

The eighteen models investigated by Tschernitz et al. are expressed in Table 4 as expectation models for y. Each temperature-dependent coefficient $k_i(T)$ or $K_i(T)$ denotes a modified Arrhenius function, centered at the mean $1/T$ value of 1/538.9 with T in Kelvins. The function form $\exp[\theta_{iA} + \theta_{iB}(1/538.9 - 1/T)]$ was tried initially to keep $k_i(T)$ and $K_i(T)$ nonnegative. This form often needed extreme parameter values to represent negligible functions; therefore, the form $\theta_{iA}\exp[\theta_{iB}(1/538.9 - 1/T)]$ with constraint $\theta_{iA} \geq 0$ was preferred.

Each weighted model $\mathfrak{F}_j$ was fitted to the weighted data Y by nonlinear least squares, using the program GREG (Stewart et al., 1992), which keeps each parameter within its permitted range. To test each model with the temperature functions included, the data for all temperatures were analyzed at once as advocated by Blakemore and Hoerl (1963).

To estimate the experimental variance, a 30-parameter polynomial in the four independent variables (T, p_H, p_U, p_S) was constructed and fitted to the data in the same manner. Reduced versions of this polynomial were then selected and tested until a least value of the residual mean square, $\hat{S}/(n - p_j)$, was found with 23 terms, giving the estimate $\nu_e = 40 - 23 = 17$ for the "pure error" degrees of freedom. The resulting weighted residual sum of squares, $\hat{S} = 60.9$, correspondingly approximates the pure error sum of squares S_e.

Our tests of the first eighteen models are summarized in Table 5, with the posterior probabilities $\pi(M_j \mid Y, S_e, \nu_e)$ calculated from Eq. 17 and normalized to a total of 1. These probabilities indicate a strong preference for Model 8, with Models 7 and 4 next best and with negligible probabilities for the other models. Model 8 also gives the best fit; its variance ratio F_j of 1.9 with the indicated degrees of freedom is exceeded with probability 0.1 in sampling from Normal error distributions, while the next best models (4 and 7) give probabilities only half as large.

Our choice of model differs from that of Tschernitz et al. (1946), who reported that Eq. 4 fitted best. The difference lies in the weightings used. Tschernitz et al. transformed each model to get a linear least-squares problem (a necessity 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 3. Analyses of Variance for Example 2

ANOVA for Model 1: Source of Variance	Sum of Squares	Deg. of Freedom	Mean Square
Residual	0.114243	34	
Lack of fit	0.070335	16	$s_1^2 = 0.004396$
Pure error	0.043908	18	$s^2 = 0.002439$
		$F_1 = s_1^2/s^2 = 1.80$	
		$Q(1.80 \mid 16,18) = 0.115$	
ANOVA for Model 2: Source of Variance	**Sum of Squares**	**Deg. of Freedom**	**Mean Square**
Residual	0.174806	33	
Lack of fit	0.130898	15	$s_2^2 = 0.008727$
Pure error	0.043908	18	$s^2 = 0.002439$
		$F_2 = s_2^2/s^2 = 3.58$	
		$Q(3.58 \mid 15,18) = 0.006$	

Table 4. Expectation Models for Response $y = \ln \mathcal{R}$ in Example 3

$f_1(\xi, \theta_1) = \ln[k_1(T)p_H/(1 + K_U(T)p_U + K_S(T)p_S)]$
$f_2(\xi, \theta_2) = \ln[k_2(T)p_U/(1 + K_H(T)p_H + K_S p_S)]$
$f_3(\xi, \theta_3) = \ln[k_3(T)p_H p_U/(1 + K_H(T)p_H + K_U(T)p_U)]$
$f_4(\xi, \theta_4) = \ln[k_4(T)p_H p_U/(1 + K_H(T)p_H + K_U(T)p_U + K_S(T)p_S)^2]$
$f_5(\xi, \theta_5) = \ln[k_5(T)p_H/(1 + K_U(T)p_U + K_S(T)p_S)^2]$
$f_6(\xi, \theta_6) = \ln[k_6(T)p_U/(1 + \sqrt{K_H(T)p_H} + K_S(T)p_S)]$
$f_7(\xi, \theta_7) = \ln[k_7(T)p_H p_U/(1 + \sqrt{K_H(T)p_H} + K_U(T)p_U)]$
$f_8(\xi, \theta_8) = \ln[k_8(T)p_H p_U/(1 + \sqrt{K_H(T)p_H} + K_U(T)p_U + K_S(T)p_S)^3]$
$f_9(\xi, \theta_9) = \ln[k_9(T)p_H/(1 + K_S(T)p_S)]$
$f_{10}(\xi, \theta_{10}) = \ln[k_{10}(T)p_H p_U/(1 + K_H(T)p_H)]$
$f_{11}(\xi, \theta_{11}) = \ln[k_{11}(T)p_H p_U/(1 + K_H(T)p_H + K_S(T)p_S)]$
$f_{12}(\xi, \theta_{12}) = \ln[k_{12}(T)p_H/(1 + K_S(T)p_S)^2]$
$f_{13}(\xi, \theta_{13}) = \ln[k_{13}(T)p_H p_U/(1 + \sqrt{K_H(T)p_H})]$
$f_{14}(\xi, \theta_{14}) = \ln[k_{14}(T)p_H p_U/(1 + \sqrt{K_H(T)p_H} + K_S(T)p_S)^2]$
$f_{15}(\xi, \theta_{15}) = \ln[k_{15}(T)p_H p_U/(1 + K_U(T)p_U + K_S(T)p_S)]$
$f_{16}(\xi, \theta_{16}) = \ln[k_{16}(T)p_H p_U/(1 + K_U(T)p_U)]$
$f_{17}(\xi, \theta_{17}) = \ln[k_{17}(T)p_U/(1 + K_S(T)p_S)]$
$f_{18}(\xi, \theta_{18}) = \ln[k_{18}(T)p_H p_U]$

Adapted from Tschernitz et al. (1946).

model according to the variance expression in Eq. 19 for ln $\mathcal{R}$. The residual sums of squares thus found are comparable to those in Table 5, and so confirm the superiority of Model 8 among those tested.

Conclusions

The main results of this work are Eqs. 12 and 17, which allow comparison of the posterior probabilities of candidate models applied to a given data set. These equations were found previously by Box and Henson (1969, 1970), but with $p(\boldsymbol{\theta}_j \mid M_j)$ regarded as an expectation before the taking of any data, a view contradicted by Eq. 11 and questioned by Kanemasu (1973). This difficulty is resolved here by evaluating $p(\boldsymbol{\theta}_j \mid M_j)$ *after* the data are fitted, as an expectation over a sample space of replicate observations generated by Eq. 3 with $\boldsymbol{\theta}_j = \hat{\boldsymbol{\theta}}_j$ and errors distributed as $N(0, \sigma^2)$. By this method, we find that Eqs. 12 and 17 hold directly for sequential investigations 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 χ^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 $2^{-p_j/2}$ in Eqs. 12 and 17 comes from the calculation of $p(\boldsymbol{\theta}_j \mid M_j)$ 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 ν_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 ν_e may be obtainable by high-order polynomial regression of the observations, or of the residuals for the best-fitting model.

Table 5. Testing of Kinetic Models*

Model	Residual Sum of Squares $\hat{S}_j$	Posterior Probability $\pi(M_j \mid Y, S_e, \nu_e)$	Var. Ratio F_j	Lack-of-Fit Deg. of Freedom $n - p_j - \nu_e$	Exp. Error Deg. of Freedom ν_e
1	970.2	0.000	13.4	19	17
2	2,156.7	0.000	29.3	20	17
3	279.5	0.014	3.6	17	17
4	192.2	0.167	2.4	15	17
5	1,013.8	0.000	14.0	19	17
6	2,586.3	0.000	33.6	21	17
7	211.1	0.213	2.3	18	17
8	165.2	0.605	1.9	15	17
9	970.2	0.000	13.4	19	17
10	844.9	0.000	11.5	19	17
11	826.2	0.000	11.9	18	17
12	1,013.8	0.000	14.0	19	17
13	767.5	0.000	10.4	19	17
14	788.8	0.000	11.3	18	17
15	420.1	0.000	5.9	17	17
16	485.1	0.000	6.2	19	17
17	2,156.7	0.000	29.3	20	17
18	925.4	0.000	11.5	21	17

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

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

$p(M_j)$ = prior probability assigned to model M_j
$p(\chi^2 \mid \nu)$ = density of χ^2 distribution with ν degrees of freedom
$s_j^2 = S_j'/(n - p_j - \nu_e)$, lack-of-fit variance for model M_j
u = observation number
$\hat{\boldsymbol{X}}_j$ = matrix of parametric sensitivities for model M_j; see Eq. 7b
$\mid \hat{\boldsymbol{X}}_j^T \hat{\boldsymbol{X}}_j \mid$ = determinant of matrix $\hat{\boldsymbol{X}}_j^T \hat{\boldsymbol{X}}_j$
$\boldsymbol{Y}$ = vector of weighted observations
$\Gamma(x)$ = gamma function at x
σ^2 = variance of observations of unit weight
$\sigma_{u_T}^2$ = expected variance of observations at ξ_u
^{T} = transpose of vector or matrix
$\hat{}$ = least-squares value
$:=$ = defines the preceding symbol by the next expression

Literature Cited

Abramowitz, M., and I. A. Stegun, eds., *Handbook of Mathematical Functions*, U.S. Government Printing Office, Washington, DC (1964); reprinted by Dover, New York (1972).

Akaike, H., "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, **AC-19**, 716 (1974).

Bates, D. M., and D. G. Watts, *Nonlinear Regression Analysis*, Wiley, New York (1988).

Bayes, T. R., "An Essay Towards Solving a Problem in the Doctrine of Chances," *Phil. Trans. Roy. Soc. London*, **53**, 370 (1763); reprinted in *Biometrika*, **45**, 293 (1958).

Blakemore, J. W., and A. W. Hoerl, "Fitting Non-Linear Rate Equations to Data," *AIChE Symp. Ser.*, **59**, 14 (1963).

Box, G. E. P., "The Exploration and Exploitation of Response Surfaces: Some General Considerations and Examples," *Biometrics*, **10**, 16 (1954).

Box, G. E. P., "Fitting Empirical Data," *Ann. N.Y. Acad. Sci.*, **86**, 792 (1960).

Box, G. E. P., "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *J. Roy. Stat. Soc.*, **A143**, 383 (1980).

Box, G. E. P., and G. A. Coutie, "Application of Digital Computers in the Exploration of Functional Relationships," *Proc. IEE*, **103**, Part B, Suppl. 1, 100 (1956).

Box, G. E. P., and D. R. Cox, "An Analysis of Transformations," *J. Roy. Stat. Soc.*, **B26**, 211 (1964).

Box, G. E. P., and T. L. Henson, "Model Fitting and Discrimination," *Dept. of Statistics Tech. Rep.*, **211**, Univ. of Wisconsin-Madison (1969).

Box, G. E. P., and T. L. Henson, "Some Aspects of Mathematical Modeling in Chemical Engineering," *Proc. Inaugural Conf. of the Scientific Computation Centre and the Institute of Statistical Studies and Research*, Cairo Univ. Press, Cairo, p. 548 (1970).

Box, G. E. P., and W. J. Hill, "Discrimination among Mechanistic Models," *Technometrics*, **9**, 57 (1967).

Box, G. E. P., and J. S. Hunter, "Multifactor Experimental Designs for Exploring Response Surfaces," *Ann. Math. Stat.*, **28**, 195 (1957).

Box, G. E. P., and W. G. Hunter, "A Useful Method for Model-Building," *Technometrics*, **4**, 301 (1962).

Box, G. E. P., and W. G. Hunter, "The Experimental Study of Physical Mechanisms," *Technometrics*, **7**, 23 (1965).

Box, G. E. P., and H. L. Lucas, "Design of Experiments in Non-Linear Situations," *Biometrika*, **46**, 77 (1959).

Box, G. E. P., and G. C. Tiao, "Multiparameter Problems from a Bayesian Point of View," *Ann. Math. Stat.*, **36**, 1468 (1965).

Box, G. E. P., and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA (1973); reprinted by Wiley, New York (1992).

Box, G. E. P., and K. B. Wilson, "On the Experimental Attainment of Optimal Conditions," *J. Roy. Stat. Soc.*, **B13**, 1 (1951).

Box, G. E. P., and P. V. Youle, "The Exploration and Exploitation of Response Surfaces: An Example of the Link Between the Fitted Surface and the Basic Mechanism of the System," *Biometrics*, **11**, 287 (1955).

Chow, G. C., "A Comparison of the Information and Posterior Probability Criteria for Model Selection," *J. Econometrics*, **16**, 21 (1981).

Daniel, C., and F. S. Wood, *Fitting Equations to Data*, 2nd ed., Wiley, New York (1980).

Gauss, C. F., *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, Perthas et Besser, Hamburg (1809); *Werke*, **7**, 240, Trans. by C. H. Davis as *Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, Little and Brown, Boston (1857); Dover, New York (1963).

Gorman, J., and R. J. Toman, "Selection of Variables for Fitting Equations to Data," *Technometrics*, **8**, 27 (1966).

Hill, P. D. H., "A Review of Experimental Design Procedures for Regression Model Discrimination," *Technometrics*, **20**, 15 (1978).

Hill, W. J., and W. G. Hunter, "A Review of Response Surface Methodology: A Literature Survey," *Technometrics*, **8**, 571 (1966).

Kanemasu, H., "Topics in Model Building," PhD Thesis, Univ. of Wisconsin-Madison (1973).

Lumpkin, R. E., Jr., W. D. Smith, Jr., and J. M. Douglas, "Importance of the Structure of the Kinetic Model for Catalytic Reactions," *Ind. Eng. Fundam.*, **8**, 407 (1969).

Mallows, C. L., "Choosing Variables in a Linear Regression: A Graphical Aid," Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, KS (1964).

Mallows, C. L., "Some Comments on C_p," *Technometrics*, **15**, 661 (1973).

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. T. Flannery, *Numerical Recipes in Fortran*, 2nd ed., Cambridge Univ. Press, Cambridge, England, p. 220 (1992).

Reilly, P. M., "Statistical Methods in Model Discrimination," *Can. J. Chem. Eng.*, **48**, 168 (1970).

Rippin, D. W. T., "Statistical Methods for Experimental Planning in Chemical Engineering," *Comput. Chem. Eng.*, **12**, 109 (1988).

Stewart, W. E., and S. M. Mastenbrook, Jr., "Parametric Estimation of Ventilation-Perfusion Ratio Distributions," *J. Appl. Physiol.: Respirat. Environ. Exercise Physiol.*, **55**(1), 37 (1983); Corr., **56**, No. 6 (1984).

Stewart, W. E., M. Caracotsios, and J. P. Sørensen, "Parameter Estimation from Multiresponse Data," *AIChE J.*, **38**, 641 (1992).

Tschernitz, J., S. Bornstein, R. B. Beckmann, and O. A. Hougen, "Determination of the Kinetics Mechanism of a Catalytic Reaction," *Trans. AIChE*, **42**, 883 (1946).

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