THE CURVE FITTING PROBLEM: A BAYESIAN APPROACH

PRASANTA S. BANDYOPADHAYAY[†]

Montana State University

ROBERT J. BOIK[‡]

Montana State University

PRASUN BASU[§]

University of Rochester

In the curve fitting problem two conflicting desiderata, simplicity and goodness-of-fit, pull in opposite directions. To this problem, we propose a solution that strikes a balance between simplicity and goodness-of-fit. Using Bayes' theorem we argue that the notion of prior probability represents a measurement of simplicity of a theory, whereas the notion of likelihood represents the theory's goodness-of-fit. We justify the use of prior probability and show how to calculate the likelihood of a family of curves. We diagnose the relationship between simplicity of a theory and its predictive accuracy.

Overview. Two conflicting desiderata, simplicity and goodness-of-fit, play key roles in fitting a curve to numerical data. Simplicity determines the shape of the curve. Goodness-of-fit, on the other hand, determines the curve that best captures the data. Working with a straight line is easy for predicting future data because it is simple. A linear equation, however, does not necessarily fit the available data; a non-linear equation may fit the data better, although the nonlinear equation is more complex. In the curve fitting problem, these two desiderata, simplicity and goodness-of-fit, pull in opposite directions. How can we make the best trade-off between these conflicting desiderata? Glymour (1980) writes, "The only moral I propose to draw (in case of the curve fitting problem) is that there is no satisfactory rationale for curve fitting available to use it." In response to Glymour, Turney (1990) has suggested a method to ease the tension between simplicity and goodness-of-fit. Forster and Sober (1994) have proposed a non-Bayesian solution to this problem.

We suggest a solution to the curve fitting problem by using Bayes' theorem. Bayes' theorem states that one can get a posterior probability if one knows the prior probability and the likelihood function. That is, $Pr(H|E) = Pr(H)Pr(E|H)/Pr(E)$, where $Pr(E) > 0$ is the marginal probability of the evidence; Pr(H) is the agent's prior probability about the hypothesis, H, before any evidence is known; and Pr(E|H) is the probability of the evidence given the hypothesis and is called the likelihood function.

In our proposal, prior probability measures the simplicity of a hypothesis. A

[†]Department of Philosophy, Montana State University, Bozeman, MT 59717.
[‡]Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717.
[§]Simon School of Business and Management, University of Rochester, Rochester, NY 14627.

Philosophy of Science, 63 (Proceedings) pp. S264–S272. 0031-8248/96/63supp-0031$2.00

hypothesis gets a higher probability than its competitors, *ceteris paribus,* if it has fewer parameters. In contrast, we say that the likelihood function measures the goodness-of-fit. A hypothesis with more parameters generally has a higher likelihood than one with fewer parameters. Given prior probability and likelihood function of a hypothesis, we get its posterior probability. We choose the hypothesis that has the highest posterior probability as making the best trade-off between simplicity and goodness-of-fit.

1. Sketch of Solution. Consider three hypotheses, H_1, H_2, and H_3 in a domain in which each is mutually exclusive of the others: H_1: $E(Y|x) = \alpha_0 + \alpha_1 x$; H_2: $E(Y|x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$; and H_3: $E(Y|x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$; where Y is a random variable, E(Y|x) is the conditional expectation of Y given x, and x is an explanatory variable. To say that these hypotheses are mutually exclusive is to say that the coefficient of x^i under hypothesis H_i is not equal to zero.

According to Bayes' theorem, the posterior probability of a hypothesis is directly proportional to prior probability of the hypothesis multiplied by its likelihood. In our view, the prior probability represents a measure of the simplicity of the hypothesis. We assign prior probability as a decreasing function of the number of parameters. That is, H_1, with the fewest parameters, gets the highest prior probability; H_2, the second highest; and H_3 and any other hypotheses with more parameters are assigned lower probabilities. Specifically, for our three hypotheses, H_1 is assigned 1/2, followed by H_2 with 1/4 and H_3 with 1/8. The remaining hypothesis, denoted by H^c, we call the *catch all* hypothesis and assign it probability 1/8. The expression "catch all" hypothesis is due to Shimony (1993). For a Bayesian account of catch-all hypothesis and its relation to the logical omniscient condition, see Earman 1992. Because the prior probability of H^c is small and we do not have any clue as to how to calculate its likelihood, we don't consider it to be a serious contender. A justification for preferring simpler hypotheses is given in Section 2.

The likelihood of the i^{th} hypothesis, denoted by L_i, provides an answer to the question how likely is the evidence given the hypothesis. The likelihood function is sometimes expressed as a probability function, $L_i = \Pr(\text{data}|H_i)$; i.e., the probability of the data given any vector of parameters $\alpha_i = (\alpha_0\alpha_1 \ldots \alpha_i)$ belonging to H_i. We choose to evaluate the likelihood function at the value of $\boldsymbol{\alpha}_i$ which makes the data most probable under H_i. That is, we take the maximum of $\Pr(\text{data}|H_i)$ over the entire parameter space of H_i and denote its value by $\hat{L}_i$. This maximal value, $\hat{L}_i$, is obtained by equating the vector of coefficients to the maximum likelihood estimate (MLE), $\hat{\alpha}_i$ and is a measure of the highest degree of support that the data can provide under a particular hypothesis. A Bayesian justification for using $\hat{L}_i$ to measure the likelihood is given in section 3.

2. Justification of Prior Beliefs. An agent's prior belief for a theory represents that agent's belief in the hypothesis before any evidence is known. Another agent may have a different prior probability for the same hypothesis. Bayesianism allows two agents to start with non-extreme divergent priors, provided their assignments of priors are consistent with the probability calculus. For this reason, Bayesians are sometimes branded as subjectivists. Bayesians not only apply Bayes' theorem, but also they interpret Bayes' theorem. According to Bayesians, as evidence accumulates, two agents with two non-extreme divergent priors that obey the probability calculus will eventually converge to strong belief in the correct hypothesis subject to certain constraints. If your degrees of belief disobey the probability calculus, then, according to Bayesians, you will be dutched provided you are involved in a bet with a Dutch bookie. For different ramifications of the Dutch-bookie argument, see Skyrms 1990. For a criticism of Bayesians, see Kyburg 1992.

Bayesians can assign any prior probability to a hypothesis, provided that the probability calculus is satisfied. Why, then, do we choose to give the highest prior probability 1/2 to the hypothesis which has fewest parameters? An answer to this question depends on our account of simplicity and its relationship to the assignment of priors. In our account, simplicity of a theory determines its prior probability. Interplay of two factors, formal and non-formal, on other hand, determines simplicity of a theory. The formal factor that is relevant to the simplicity consideration is paucity of parameters. The non-formal factors that play key roles in determining simplicity are epistemological and pragmatic factors.

We consider H_1 as the simplest hypothesis because we find that it is easiest to work with a hypothesis with fewer parameters. Though our selection of H_1 as the simplest hypothesis is based on a pragmatic consideration, this pragmatic consideration is not necessarily devoid of any relationship with our epistemic reason for embracing H_1 as the simplest hypothesis.

Many philosophers, including van Fraassen (1980), contend that reasons for accepting a hypothesis may be numerous. Some reasons for acceptance are pragmatic, whereas others are epistemic. Van Fraassen thinks that epistemic reasons for acceptance of a theory cannot be pragmatic reasons for acceptance of a theory and vice versa. In contrast, Harman (forthcoming) convincingly argues that the distinction between pragmatic reasons and epistemic reasons need not be exclusive. In other words, pragmatic reasons can sometimes be epistemic reasons and vice versa. Following Harman, we contend that if a consideration makes a difference in the probability of embracing a theory, then it is an epistemic reason for embracing the theory. If there is a pragmatic consideration, in the light of which we decide that the simplest theory has the highest prior probability, then this is a pragmatic-epistemic reason for believing the theory. That is, it is likely that a simpler theory will be true.

3. Bayesian Justification for Maximizing the Likelihood. In the example of section 1, we assigned higher likelihoods to hypotheses which had many adjustable parameters. The pertinent question, however, is how can we calculate the likelihood of a family of curves? By a family of curves, we mean the infinite set of curves generated by allowing the coefficients $\alpha_0, \ldots\ldots, \alpha_i$ to take on any values subject to $\alpha_i \neq 0$. Our proposal is to calculate the likelihood of a family of curves by the likelihood of *the best fitting curve* in that family. Bennett (in private discussion) urges on us the need for focusing on the likelihood of the best fitting curve while calculating the likelihood of a family of curve.

One criticism of the proposed approach is that it appears to assign prior probability 1 to the maximum likelihood estimator, $\hat{\alpha}_i$. Forster while commenting on a previous version of this paper (presented at the APA meeting, Central Division, 1995), has raised this objection. Sober (in private correspondence) has also objected to the same point. Because the MLE cannot be computed until the data are observed, our prior *appears* to depend on the data. In this section, we show that this is not the case. We do not assign prior probability of 1 to the MLE. On the contrary, our approach spreads the prior probability for the vector α_i over the whole of i + 1 dimensional space. The MLE, $\hat{\alpha}_i$, is but one point in this space. Our approach is identical to what Rosenkrantz (1977) called the method of averaged likelihoods, but our choice of prior distributions for the unknown parameters differs from his.

Suppose that there are k hypotheses (k = 3 in our example), where each hypothesis corresponds to a family of polynomial curves. The i^{th} hypothesis can be written as H_i: $\mathrm{E(Y|x)} = \sum_{j=0}^{\mathrm{i}} \alpha_j \mathrm{x}^{\mathrm{j}}$ where α_j for $j = 1, \ldots, i$ are unknown regression coefficients and x is a known *explanatory* variable. A sample of *n* data points (i.e., the evidence) will be drawn. Denote the sample by $(\mathbf{Y}, \mathbf{x})$, where $\mathbf{Y} = (\mathrm{Y}_1\ \mathrm{Y}_2 \ldots \mathrm{Y}_n)'$ and $\mathbf{x} = (\mathrm{x}_1\ \mathrm{x}_2 \ldots \mathrm{x}_n)'$. It is assumed that if H_i is true, then the data points are independently and normally distributed with mean $\mathrm{E}(\mathrm{Y}_t|x_t) = \sum_{j=0}^{\mathrm{i}} \alpha_{\mathrm{j}} \mathrm{x}_{\mathrm{t}}^{\mathrm{j}}$ and variance σ^2. This normality assumption is conveniently summarized as

$$Y|\mathrm{H}_i, \boldsymbol{X}_i, \boldsymbol{\alpha}_i, \sigma^2 \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\alpha}_i, \sigma^2\mathbf{I}_n), \tag{1}$$

where $\mathbf{X}_{\mathrm{i}}$: n × (i + 1) and $\boldsymbol{\alpha}_{\mathrm{i}}$: (i + 1) × 1 are given by

$$\mathbf{X}_i = \begin{pmatrix} 1 & \mathrm{x}_1 & \mathrm{x}_1^2 & \cdots & \mathrm{x}_1^{\mathrm{i}} \\ 1 & \mathrm{x}_2 & \mathrm{x}_2^2 & \cdots & \mathrm{x}_2^{\mathrm{i}} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \mathrm{x}_{\mathrm{n}} & \mathrm{x}_{\mathrm{n}}^2 & \cdots & \mathrm{x}_{\mathrm{n}}^{\mathrm{i}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\alpha}_{\mathrm{i}} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{\mathrm{i}} \end{pmatrix}.$$

To compute the posterior distribution of H_i, prior distributions on H_i and on the parameters $\boldsymbol{\alpha}_i$ and σ^2 must be specified. In Section 1, prior probabilities were assigned to H_i using an inverse relationship: the fewer the number of adjustable parameters in H_i, the larger its prior probability. This prior can be stated as

$$\Pr(\mathrm{H}_i) = \theta \quad \text{for} \quad i = 1, \ldots, k. \tag{2}$$

For the example of Section 1, $\theta_1 = 1/2$, $\theta_2 = 1/4$, and $\theta_3 = 1/8$. The remaining hypotheses were collectively assigned probability 1/8.

For the assignment of priors to $\boldsymbol{\alpha}_i$ and σ^2, it is assumed that little is known about these parameters before collecting data. Accordingly, diffuse priors will be adopted. Diffuse priors spread the probability over the entire parameter space in such a way that no points are greatly favored over others. Specifically, we assume that

$$\boldsymbol{\alpha}_i|\mathbf{X}_i, \sigma^2, \tau, \bar{\alpha}_{\mathrm{i}} \sim \mathrm{N}(\bar{\boldsymbol{\alpha}}_{\mathrm{i}}, \mathbf{V}_i); \text{ and that } \ln(\sigma^2) \sim \mathrm{Uniform}(-\infty, \infty); \tag{3}$$

where $\bar{\boldsymbol{\alpha}}_{\mathrm{i}}$ is the prior mean of $\boldsymbol{\alpha}_i$; $\mathbf{V}_i$ is the prior variance of $\boldsymbol{\alpha}_{\mathrm{i}}$; $\mathbf{V}_i = \boldsymbol{\sigma}^2\, \tau^{1/(\mathrm{i}+1)}(\mathbf{X}_i'\mathbf{X}_i)^{-1}$ and τ is a large positive constant. The prior mean of $\boldsymbol{\alpha}_i$ can be assigned any value because the posterior distribution of the hypothesis depends on $\bar{\boldsymbol{\alpha}}_{\mathrm{i}}$ only minimally when τ is large.

The prior adopted for $\boldsymbol{\alpha}_i$ is a conjugate prior (Berger 1985) for the normal density of $\mathbf{Y}$ in (1). Furthermore, the prior on $\boldsymbol{\alpha}_i$ is a special case of the prior used by Smith and Spiegelhalter (1980) as well as a special case of the g-prior suggested by Zellner (1986). Further details concerning the prior on $\boldsymbol{\alpha}_i$ are available from the authors. The parameter τ controls how diffusely the prior probability on $\boldsymbol{\alpha}_i$ is spread over the $i + 1$ dimensional space. As the value of τ increases, the prior distribution on $\boldsymbol{\alpha}_i$ becomes more diffuse. In our calculation of posterior probabilities, we will take τ to be an infinitely large value. The prior on σ^2 is improper and was suggested by Jeffreys (1961). It is an invariant diffuse prior and says that on the log scale, σ^2 is equally likely to be in any interval of fixed size.

The posterior probability of H_i conditional on the data is given by Bayes' theorem as

$$\Pr(\mathrm{H}_i|\mathbf{Y}, \tau, \bar{q}, \ldots, \bar{\boldsymbol{\alpha}}_{\mathrm{k}}) = \Pr(\mathbf{Y}|\mathrm{H}_i\tau, \bar{\alpha}_1, \ldots, \bar{\alpha}_{\mathrm{k}}) \Pr(\mathrm{H}_i)/\Pr(\mathbf{Y}|\tau, \bar{\alpha}_1, \ldots, \bar{\alpha}_{\mathrm{k}}).$$

Using the probability model in (1) and the priors in (2) and (3), the posterior probability, for fixed τ and $\bar{\boldsymbol{\alpha}}_1, \ldots, \bar{\boldsymbol{\alpha}}_k$, is

$$\Pr(\mathrm{H}_i|\mathbf{Y}, \tau, \bar{\boldsymbol{\alpha}}_i, \ldots, \bar{\boldsymbol{\alpha}}_k)$$

$$= \frac{\Pr(H_i) \int \Pr(\mathbf{Y}|\mathbf{X}_i, \boldsymbol{\alpha}_i, \sigma^2) \Pr(\boldsymbol{\alpha}_i|\mathbf{X}_i, \sigma^2, \tau, \bar{a}_i) \Pr(\sigma^2) d\boldsymbol{\alpha}_i d\sigma^2}{\sum_{j=1}^{k} \Pr(\mathrm{H}_j) \int \Pr(\mathbf{Y}|\mathbf{X}_j, \sigma_j, \sigma^2) \Pr(\boldsymbol{\alpha}_j|\mathbf{X}_j, \sigma^2, \tau, \bar{\boldsymbol{\alpha}}_j) \Pr(\sigma^2) d\boldsymbol{\alpha}_j d\sigma^2}$$

$$= \frac{\theta_i[SSE_i + (\hat{\boldsymbol{\alpha}}_i - \bar{\boldsymbol{\alpha}}_i)'\mathbf{X}_i'\mathbf{X}_i(\hat{\boldsymbol{\alpha}}_i - \bar{\boldsymbol{\alpha}}_i)]^{-n/2}\varphi_i^{-(i+1)/2}}{\sum_{j=1}^{k} \theta_j[SSE_j + (\hat{\boldsymbol{\alpha}}_j - \bar{\boldsymbol{\alpha}}_j)'\mathbf{X}_j'\mathbf{X}_j(\hat{\boldsymbol{\alpha}}_j - \bar{\boldsymbol{\alpha}}_j)\varphi_j^{-1}]^{-n/2} \varphi_j^{-(j+1)/2}},$$

where $\varphi_j = 1 + \tau^{1/(j+1)}$ for $j = 1, \ldots, k$;

$$\hat{\boldsymbol{\alpha}}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Y}; \text{ and } \mathrm{SSE}_i = (\mathbf{Y} - \mathbf{X}_i\hat{\boldsymbol{\alpha}}_i)'(\mathbf{Y} - \mathbf{X}_i\hat{\boldsymbol{\alpha}}_i). \tag{4}$$

The quantities in (4) are the maximum likelihood estimator of $\boldsymbol{\alpha}_i$ under hypothesis H_i and the corresponding sum of squared residuals (lack of fit measure) from the maximum likelihood fit, respectively. Mathematical details concerning the required integrations can be obtained from the authors.

Our approach is to compute $\Pr(\mathrm{H}_i|\mathbf{Y})$ as the limiting probability as $\tau \to \infty$. The result is $\Pr(\mathrm{H}_i|\mathbf{Y})$

$$= \lim_{\tau \to \infty} \Pr(H_i|\mathbf{Y}, \tau, \bar{\boldsymbol{\alpha}}_1, \ldots, \bar{\boldsymbol{\alpha}}_k) = \theta_i SSE_i^{-n/2} \Big/ \left[\sum_{j=1}^{k} \theta_j SSE_j^{-n/2}\right]. \tag{5}$$

Alternatively, one can say that the posterior probability of H_i is proportional to the numerator of (5) because the denominator becomes constant when conditioning on the data. That is,

$$\Pr(\mathrm{H}_i|\mathbf{Y}) \propto \theta_i\lambda/SSE_i^{n/2}, \tag{6}$$

where λ is any constant.

In Section 1, it was proposed to compute the posterior probability by multiplying the prior probability of H_i by the maximized likelihood function. Using the notation of this section, the proposal was to compute $\Pr(\mathrm{H}_i|\mathbf{Y})$ as

$$\Pr(\mathrm{H}_i|\mathbf{Y}) \propto \theta_i\hat{\mathrm{L}}_i = \theta_i \max \exp\{-(\mathbf{Y} - \mathbf{X}_i\boldsymbol{\alpha}_i)'(\mathbf{Y} - \mathbf{X}_i\boldsymbol{\alpha}_i)/(2\sigma^2)\}/(2\pi\sigma^2)^{n/2},$$

where the maximum is taken with respect to $\boldsymbol{\alpha}_i$ and σ^2. Equating the derivatives with respect to unknown parameters to zero and solving the resulting normal equations yields (6), with $\lambda = e^{-n/2} (2\pi/n)^{-n/2}$. Thus, if one adopts the priors in (2) and (3), then the posterior probability is obtained by multiplying the prior probability of H_i by the maximized likelihood function.

4. Illustration of Likelihood Calculations. Consider the following example. Suppose that Sue heats her house with natural gas. The amount of gas required to heat the home depends on the outside temperature. That is, if the weather is cold, Sue needs more gas to heat her house as long as her family's habits, the insulation of the house and the other relevant factors remain unchanged. She measures her household's natural gas consumption each month during one heating season, from October to the following June. For the sake of simplicity, we assume each month consists of 30 days. Outside temperature influences gas consumption only when it is cold enough to require heating. We measure the usual need for heating in degree days. One heating degree day is accumulated for each degree the average daily temperature falls below 65 degree Fahrenheit. An average temperature of 20°F, for example, corresponds to 45 degree days. In Table I, the explanatory variable, x, is heating degree days per day for the month, and the response variable, Y, is gas consumption per day in hundreds of cubic feet

TABLE I

	Month								
Variable	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
x	15.6	26.8	37.4	36.4	35.5	18.6	15.3	7.9	0.0
Y	5.2	6.1	8.7	8.5	8.8	4.9	4.5	2.5	1.1

The following summary statistics were calculated using equations (4) and (6).

TABLE II

	Maximum Likelihood Estimates						
Hypothesis	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	*SSE*	$\hat{L}$	Pr(H|Y)
1	1.221	0.203			1.300	0.0173	0.0087
2	1.095	0.223	−0.0005		1.259	0.0199	0.0050
3	0.975	0.300	−0.0064	0.0001	1.140	0.0310	0.0039

Table II shows that H_1 makes a better trade-off between simplicity and goodness-of-fit than either H_2 or H_3. Even though the likelihoods of H_2 and H_3 are larger than that of H_1, the posterior probability for H_1 is largest because of the impact of higher prior probability assigned to H_1. In the fifth and final section, we will discuss the implications of this example.

5. Predictive Accuracy and Simplicity. Predictive accuracy is a comparative concept. We say that one theory has a greater predictive accuracy than another only if the former is much closer to the truth than the latter. The evidential role of the simplicity of a theory is related to its predictive accuracy. The evidential role simplicity plays can be understood in two ways: (i) We can evaluate the simplicity of a hypothesis with respect to its retrodiction and (ii) we can consider a theory's predictive accuracy in the future.

Philosophers disagree as to whether simplicity has an evidential role in theory choice. Different stripes of instrumentalists contend that simplicity *never* plays an evidential role. According to them, simplicity is a pragmatic reason for embracing a theory. Realists, on the other hand, argue that simplicity *sometimes* plays an evidential role in theory appraisal. Quine (1992) and van Fraassen belong to the former group, whereas Forster and Sober belong to the latter. We think both are mistaken in taking simplicity to be playing one role or another in theory choice. Both instrumentalists and realists are mistaken, for whether a simple theory has better predictive accuracy depends on what future data we are confronted with. If the future data are more amenable to a simpler hypothesis like H_1, then the simpler hypothesis is likely to have a better predictive accuracy than a less simple hypothesis, and so to have an evidential role to play in theory appraisal. If, however, the new data fit well with a complex hypothesis H_2, then it has better predictive accuracy and so has an evidential role to play. The key idea in understanding the connection, if any, between simplicity of a theory and its predictive accuracy is related to what kind of future data, linear or quadratic, we confront.

Consider the posterior probabilities in Table II. Hypothesis 1 looks to be most predictively accurate, given currently available evidence. When we take into account some future data, however, then we might find that a less simple hypothesis, say H_2, gives better predictive accuracy. Recall the linear equation $E(Y|x) = \alpha_0 + \alpha_1 x$ under H_1. Sue wants to know what her gas consumption will be next February. Suppose that next February, her gas consumption is 870 cubic feet per day. We cannot compare this year's gas consumption with last February rate (880 cubic feet per day) unless the average temperatures for the two months are the same. Suppose that next February has an average of 40 degree days. We therefore *forecast* from the regression equation how much gas the house would have used at 40 degree day this year. Our forecast is $\hat{Y} = 1.221 + (0.203)\,40 = 9.341$ or 934 cubic feet. Sue estimates that she saved about 64 cubic feet per day.

Compare this prediction with the prediction based on H_2: $\hat{Y} = 1.095 + 0.223x - 0.0005x^2 = 9.243$. In other words, 924 cubic feet per day. This gives a more accurate prediction than that based on H_1. Though the prediction does not agree with the exact amount of gas consumption during the month of February, it is much closer to the actual value than that of made on the basis of H_1.

This example shows the senses in which the predictive accuracy of a theory does not always depend on the simplicity of a theory. Although H_1 is simpler than H_2, H_2 is predictively more accurate than H_1, at least for next February. One possible pragmatist's contention is that any linear hypothesis is always simpler than any other nonlinear hypothesis. This fact is independent of whether the theory in question is predictively accurate. If two theories are predictively accurate, then according to pragmatists, we can choose the simpler of the two theories. If theories are not predictively accurate, then, if simplicity is our sole concern, we can go for the simpler theory. In both the cases, for pragmatists, simplicity of a theory has nothing to do with its predictive accuracy and therefore, has no evidential role to play in theory appraisal.

Our response to this defense of simplicity is that simplicity is not the sole criterion in theory choice. Informativeness or empirical adequacy of a theory plays an equally important role in preferring one theory to the other. Pragmatists like van Fraassen, however, won't agree with this pragmatist's rejoinder to our argument. For him, the necessary condition for acceptance of a theory involves the belief that the theory in question is empirically adequate. In the above scenario, the linear hypothesis is not an empirically adequate theory, since it does not provide a predictively accurate account of the relation between the response variable and the explanatory variable. Since the linear hypothesis is not an empirically adequate theory, van Fraassen won't consider it to be a contending hypothesis. We do not know Quine's position at this point.

To see why the nature of future data is crucial in this situation, we will enrich this example with further details. The next February, Sue's gas consumption is, in fact, 870 cubic feet. Our additional information is that Sue adds insulation to her attic during the summer, anticipating that her gas consumption will be reduced in the coming year. Given this information, Sue saved about 64 cubic feet per day by adding insulation based on the prediction of the linear hypothesis. In contrast, based on the prediction of the second degree equation, Sue estimates saving 54 cubic feet per day.

If we furnish our situation with this additional information, Forster and Sober may argue, that this does not show that the predictive accuracy of the second degree hypothesis is better than the first degree hypothesis. In their rejoinder, they seem to be saying that this further information that Sue added insulation to her attic during the summer, is not what the linear hypothesis is supposed to capture at the time of forecasting. When these two hypotheses are constructed, they are constructed based on information available at that time. Since the added information is not available at the time of the construction of these hypotheses, the claim that the first degree hypotheses is not predictively accurate, according to them, changes the subject if we conclude from this that the simpler of the two hypotheses fails to be predictively accurate. I think that they are right in pointing out that the new information is not what we can expect of a linear or nonlinear hypothesis to predict beforehand.

Our discussion about the Sue example shows that a simpler theory has an evidential role to play in theory choice (since it has a greater predictive accuracy than its complex counterparts), does not make sense unless we know what kind of future data we will be predicting based on this simpler theory. This is argued in two ways: (i) In the Sue example *before* any further information added and (ii) in the Sue example *after* any further information added.

(i) In the first case, a simpler theory will have greater predictive accuracy so long as the future data are amenable to this simple hypothesis. Even though H_1 has the highest posterior probability, as soon as we bring in new datum, H_2 or H_3 could provide more accurate predictions.

(ii) In the second case, i.e. the Sue example, after any further information is added, Forster and Sober rightly point out that we cannot expect our linear or nonlinear hypotheses to anticipate the fact about Sue that she adds insulation to her attic during the summer so that the next February her gas bill will be appreciably lower. So, it is clearly a change of subject. Recall, however, that the first degree hypothesis is not predictively accurate at any rate *before* any further information is supplied to us. In this situation, whether the first order hypothesis fails to be predictively accurate depends on what future data we encounter. In any case, to appreciate the evidential relations between simplicity of a theory and its predictive accuracy so that we can say that a simpler theory is a good predictor, we need to know the nature of future data. This is a point which Forster and Sober's recent article overlooks.

Summing Up. In the curve fitting problem two conflicting desiderata, simplicity and goodness-of-fit pull in opposite directions. To this problem, we proposed a Bayesian solution that strikes a balance between simplicity and goodness-of-fit. Using Bayes' theorem we argued that the notion of prior probability represents a measure of simplicity of a theory, whereas the notion of likelihood represents the theory's goodness-of-fit. We justified the use of prior probability and showed how to calculate the likelihood of a family of curves. We also diagnosed the relationship between simplicity of a theory and its predictive accuracy.

REFERENCES

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis,* Second edition. New York: Springer-Verlag.

Earman, J. (1992), *Bayes or Bust?.* Cambridge, MA: MIT Press.

Forster, M. and Sober, E. (1994), "How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions", *British Journal for Philosophy of Science* 45, 1–35.

Glymour, C. (1980), *Theory and Evidence.* Princeton: Princeton University Press.

Harman, G. (Forthcoming), "Pragmatism and Reasons for Belief" in C. B. Kulp (ed.), *Realism/Antirealism and Epistemology.* Totowa, NJ: Rowman and Littlefield.

Jeffreys, H. (1961), *Theory of Probability,* third edition. New York: Oxford University Press.

Kyburg. H. (1992), "The Scope of Bayesian Reasoning", in D. Hull, M. Forbes and K. Okruhlik (eds.), *PSA-1992,* 2, East Lansing: The Philosophy of Science Association, pp. 139–152.

Quine, W (1992), *The Pursuit of Truth.* Cambridge: Harvard University Press.

Rosenkrantz, R. D. (1977), *Inference, Method and Decision.* Boston: D. Reidel Publishing Company.

Shimony, A. (1993), "Scientific Inference" in *Search for a Naturalistic World View,* Vol. 1. New York: Cambridge University Press.

Skyrms, B (1990), *The Dynamics of Rational Deliberations.* Cambridge, MA: Harvard University Press.

Smith, A. F. M., & Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models", *Journal of the Royal Statistical Society,* Series B 42, 213–220.

Turney, P. (1990), "The Curve Fitting Problem—A Solution", *British Journal for the Philosophy of Science* 25, 509–530.

Van Fraassen, B. (1980), *The Scientific Image.* New York: Oxford University Press.

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions", in P. K. Goel and A. Zellner, (eds.), *Bayesian Inference and Decision Techniques.* New York: Elsevier Science Publishing Company, Inc., pp. 233–243.