NONPARAMETRIC FUNCTION-FITTING TO SUGGEST NONLINEAR PARAMETRIC MODELS

Matthias Schonlau, Michael Hamada, and William Welch, University of Waterloo
Matthias Schonlau, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1 Canada

Key Words: Nonparametric function fitting, Computer experiments, Nonlinear regression, Model identification, Graphical methods

1 Introduction

The computer models we have experienced are evaluated through complex, deterministic code. Given the same inputs, these computer models always reproduce the same outputs. Typically, the code will to be expensive to run, e.g., it may solve a large number of differential equations which may require several hours or more of computer time. Because of the code's computer intensive requirements, computer experiments have been employed for finding a good cheap-to-compute surrogate (i.e., predictor) for the computer model.

Thus far computer experiments have focused on building computationally cheap nonparametric predictors. See, for example, Currin et al. (1991), Sacks, Schiller and Welch (1989), Sacks, Welch, Mitchell and Wynn (1989), and Welch, Buck, Sacks, Wynn, Mitchell, and Morris, (1992). In recent work with some mechanical engineers, however, *explanation* rather than prediction was the overriding objective. The engineers had developed a computer code for the performance of a solar collector and wanted to know quantitatively what the impact of six design factors was on the response, heat exchange effectiveness. More specifically, they were interested in finding a *parametric* model that explained the complex functional relationships embodied in their computer code.

In general, finding a *parametric* model is nontrivial because the typical relationship is highly nonlinear and there are no obvious classes of models from which to choose. In this paper we propose identifying approximating parametric models from a graphical analysis of a nonparametric model. Note that here the nonparametric analysis of computer experiments is an intermediate tool rather than an end in itself. As will be shown, the identification process is fairly automatic.

An overview of this section is as follows. Section 2 first outlines a nonparametric method for analyzing data from a computer experiment and then explains how parametric models can be identified graphically. Section 3 demonstrates these ideas using data from a computer experiment on the solar collector code mentioned above. Some discussion regarding the choice of experimental design is also given. The section concludes with a discussion in Section 4.

2 Identifying Parametric Nonlinear Models

Identifying a class of nonlinear models that fit the data well is easy if there is only one covariate. A simple scatter plot would reveal the functional relationship which for a computer model is exact since the relationship is deterministic. Then, the data analyst can choose a class of models suggested by the scatter plot and fit the model using standard nonlinear regression software to obtain parameter estimates. This approach was used in a case study presented in Bates and Watts (1988, Section 3.13) for physical experimental data which contained random error. While the data from a computer experiment contain no random error, the objective here remains the same, i.e., to find a good approximating model.

However, scatter plots are not very useful for model identification where there is more than one covariate, as the relationship between the response and each covariate can be masked by the relationships between the response and the other covariates (e.g., Montgomery and Peck, Section 4.2.5). To overcome the masking problem, a plot of a function involving only the covariate of interest is needed. In other words, the effects of the other covariates need to be eliminated. Such plots will be considered shortly after some preliminaries.

First, a brief overview of a nonparametric predictor of the computer model is given because the

nonparametric predictor plays a key role in the method proposed shortly. The data from a computer experiment consist of n vectors of covariate values (or inputs) denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_n$ for the k covariates $x_1, \ldots, x_k$ as specified by a particular experimental design. The corresponding response values (or outputs) are denoted $\mathbf{y} = (y_1, \ldots, y_n)^t$. Then, following the approach of, e.g., Welch et al. (1992), the response is treated as a realization of a stochastic process:

$$Y(\mathbf{x}) = \beta + Z(\mathbf{x}),$$

where $E(Z(\mathbf{x})) = 0$ and $\mathrm{Cov}(Z(\mathbf{w}), Z(\mathbf{x})) = \sigma^2 R(\mathbf{w}, \mathbf{x})$ for two inputs $\mathbf{w}$ and $\mathbf{x}$. The correlation function $R(\cdot, \cdot)$ can be tuned to the data, which for this paper is assumed to have the form:

$$R(\mathbf{w}, \mathbf{x}) = \prod_{j=1}^{k} \exp(-\theta_j |w_j - x_j|^{p_j}), \qquad (1)$$

where $\theta_j \geq 0$ and $0 < p_j \leq 2$. The p_j's can be interpreted as smoothness parameters (smoother as the p's increase) which indicate the smoothness of the response surface and the θ's indicate how local the predictor is (more local as the θ's increase).

The best linear unbiased predictor of y at an untried $\mathbf{x}$ can be shown to be:

$$\hat{y}(\mathbf{x}) = \hat{\beta} + \mathbf{r}^t(\mathbf{x})\mathbf{R}^{-1}(\mathbf{y} - 1\hat{\beta}), \qquad (2)$$

where $\mathbf{r}^t(\mathbf{x})$ is the vector of the correlations between $\mathbf{x}$ and each of the n design points, $\hat{\beta}$ is the generalized least squares estimator of β, $\mathbf{R}$ is the correlation matrix with elements defined by (1) and 1 is a vector of 1's. While this cheap-to-compute predictor has proven to be accurate for numerous applications, it does not reveal the relationship between y and $x_1, \ldots, x_k$ in a readily interpretable way. Consequently, this predictor is unsuitable for *explaining* the functional relationship between the covariates and the response.

It is worth considering the situation when the functional relationship between the covariates $x_1, \ldots, x_k$ and the response y is additive:

$$y = \mu_0 + \mu_1(x_1) + \mu_2(x_2) + \ldots + \mu_p(x_k).$$

In this situation, the difficult problem of identifying a nonlinear function $y(x_1, \ldots, x_k)$ has turned into the much easier problem of identifying $\mu_i(x_i)$ for $i = 1, \ldots, k$. Note that while the method proposed later in this section does not assume additivity, an important point is that additivity does make the model identification much easier.

Recall that in order to identify the functional relationship between a group of covariates and the response, the effect of each of these covariates needs to be isolated from the remaining ones. When we want to isolate the effect of a single covariate, the true *main* effect of the covariate can be calculated in the following two ways by:

1. *Integrating out the other factors.* The main effects are defined as:

$$\mu_i(x_i) = \int y(\mathbf{x}) \prod_{h \neq i} dx_h,$$

(Sacks et al., 1989). They can be estimated by replacing $y(\mathbf{x})$ by $\hat{y}(\mathbf{x})$.

2. *Keeping the other variables fixed.* For example, the other variables might be fixed at their respective midranges.

In both calculations, the unknown $y(\mathbf{x})$ needs to be replaced by $\hat{y}(\mathbf{x})$ from Equation (2). The first approach is preferred because it is analogous to analysis of variance in that all the other covariates are averaged out. Note also that integrating $\hat{y}(\mathbf{x})$ is numerically easy to perform if the $\mathbf{x}$ region is cuboidal and if the correlations are in product from as in (1). In a similar fashion, the joint effect of two or more covariates can be investigated by integrating out all the other covariates or fixing the other covariates at specific values.

Main effects or joint effects (say, of two covariates) can then be displayed graphically for each x_i and pair (x_i, x_j). By choosing a tentative model for each of the effect plots which displays some key feature (i.e., impacts the response), an overall model can be developed by adding up all the corresponding candidate models.

If there are no interactions (and hence, additivity holds) the k-dimensional problem has been reduced to k one-dimensional problems. If large interactions are present, then the interacting covariates need to be considered jointly. Covariates can then be grouped so that covariates in two different groups do not interact. Provided that the groups contain no more than two variables, candidate models may still be identified from contour plots of the response. For larger sized groups, such plots will generally not be helpful. In this case, when faced with many interactions, transforming the response may help in reducing the apparent complexity. Experience with a number of computer models, however, suggests the complexity of computer models tends to

arise from nonlinearities rather than through interactions.

Subsequently, the identified parametric model can be fit using standard nonlinear regression techniques. When there is additivity, starting values for the parameter estimates can be estimated from the main effect plots.

3 Application to a Solar Collector Code

In this section, the proposed method is applied to an expensive-to-compute computer model for the heat exchange effectiveness between the air and an unglazed transpired-plate solar collector with slot-like perforations (henceforth, referred to as holes). The use of equally spaced slot-like holes replaces the unrealistic assumption of infinitesimally small and infinitesimally close holes and thus, represents an engineering novelty in the design of unglazed solar collectors. Golneshan (1991) showed that the heat exchange effectiveness for these solar collectors is a function of six covariates, (1) inverse wind velocity, (2) dimensionless slot width, (3) Reynolds number, (4) admittance, (5) dimensionless plate thickness, and (6) the radiative Nusselt number, as defined by a system of differential equations. The computer code (Cao, 1993) solves the system of differential equations for given covariate values and requires around two hours of computing time on a workstation. The response considered here is the increase in heat exchange effectiveness attributed to the heat transfer in the holes from the hole sides and is expressed as a percentage (0-100). For further details, see Cao (1993). For notational simplicity, in the following, the six covariates listed above will be referred to as $x_1, x_2, \ldots, x_6$ and the response as y.

The mechanical engineers who had developed the solar collector code were interested specifically in explaining the impact of the six covariates (which are design factors) on the response heat exchange effectiveness; ultimately, the explanation would help to identify better solar collector designs. Note that such understanding was not apparent from inspecting the system of differential equations. The engineers were interested in developing a surrogate *parametric* model because empirical models of this type existed in the literature for solar collectors based on older technologies; they had no preconceived idea of what form the model should take because the collectors with slot-like holes represented state-of-the-art technology. Hence, the need arose for performing an experiment on the solar collector code, i.e., a computer experiment.

The experimental design used for the computer design was one that filled the six dimensional cuboidal region, a so-called space filling design. Specifically, a Latin hypercube design (McKay, Beckman, and Conover, 1979) consisting of 100 points was chosen in which the minimum distance between points (i.e., the covariate vectors) was maximized. The design was found using ACED (Algorithms for Constructing Experimental Designs) which was developed by Welch. All the two-dimensional projections of the Latin hypercube design can be seen in Figure 1, which shows that the design is indeed space-filling.

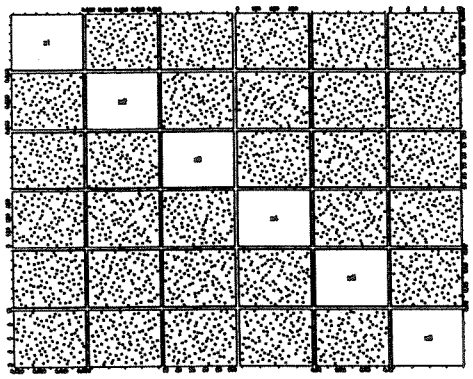

Figure 1: Two Dimensional Projections of the Latin Hypercube Design

Because the effectiveness response is a percentage, models in the logit of the response were considered. Scatter plots of the logit data (Figure 2) indicate a possible linear trend in x_2 and x_5. The remaining relationships, if any, are masked by the presence of the other covariates. In the following, the proposed method for identifying a class of surrogate nonlinear models will be applied.

The stochastic process predictor (2) for the logit response was fit using the nonparametric method outlined in the previous section by GaSP (Gaussian Stochastic Processes), also developed by Welch. Figure 3 displays a cross validation plot for the fit in which each observed y is predicted using all the remaining data and shows that the predictor is reasonably accurate. Main effect plots were then generated using Method 1 (integrating out the other covariates) as shown in Figure 4 for covariates x_1 through x_5. The main effect for covariate x_6 is very flat, and all two-way interactions are close to zero everywhere. These effects were considered negligible by the engineers. The features displayed in the

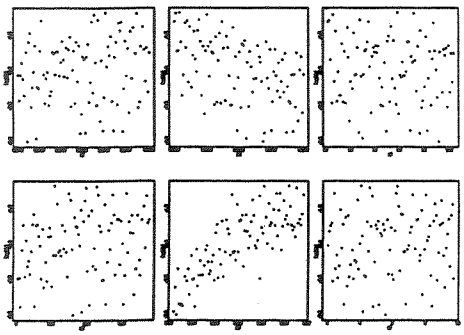

Figure 2: Scatter Plots of logit(y) versus x_i

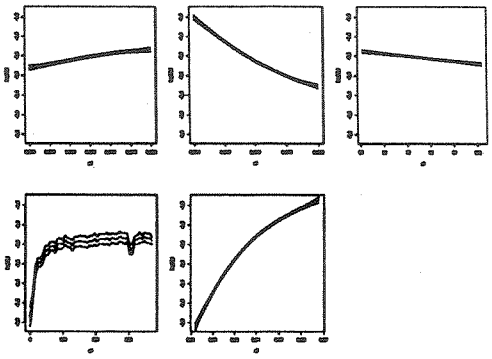

Figure 4: Main Effects for $\mathbf{x}_1$ through $\mathbf{x}_n$, and $\pm$ 2 S.E. of the Effects

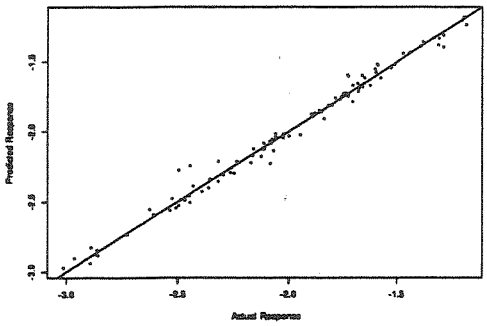

Figure 3: Cross Validation of the Nonparametric Model

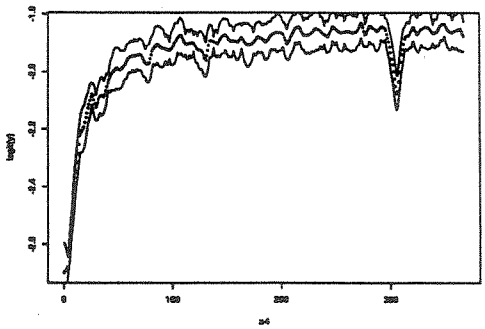

Figure 5: Predicted $\hat{y}(\mathbf{x})$ versus x_4 Using Method 2, and $\hat{y}(\mathbf{x}) \pm 2$ S.E.$[\hat{y}(\mathbf{x})]$, where all other x Variables are Fixed at their Midranges

main effect plots suggest that the effect of x_1, x_2, and x_3 are approximately linear and the effect of x_5 is approximately quadratic.

The main effect plot for x_4 is rather ragged. Although the plot gives a good indication of the apparently nonlinear x_4 effect, it is doubtful that the true x_4 relationship is that bumpy. One possible explanation is that the computer code may have some numerical convergence problems in certain regions of the x space. This possible erratic behavior may then be erroneously attributed to x_4 which clearly has the most nonlinear or complex impact on the response. Plots of the main effects using Method 2, where the other variables are fixed at their respective midranges rather than being averaged out, result in very similar graphs. For example, Figure 5 shows the Method 2 main effect plot for x_4.

The nonlinear shape of the x_4 main effect plot which appears to asymptote can be captured by a Michaelis-Menten model (Bates and Watts, 1988, p. 329); the Michaelis-Menten model has long been used to model the behavior of a limiting chemical reaction which rises at a decreasing rate to an asymptote. A Michaelis-Menten type model also arises in the context of a reciprocal link function in a generalized linear model, where an inverse linear response function is assumed (McCullagh and Nelder, 1989, p. 291). Here, to give more flexibility, the Michaelis-Menten model was augmented by introducing an additional parameter β_2 and takes the following form:

$$y = \frac{1}{\beta_0 + \beta_1 / x_4^{\beta_2}}.$$

The overall model consisting of linear effects in x_1, x_2, x_3, and x_5, a quadratic effect in x_5 and the augmented Michaelis-Menten model for x_4 was then fit using standard nonlinear regression software which gave:

$$\text{logit}(y) = 6.601x_1 - 0.0028x_3 - 35.41x_2 + 53.61x_5 - 392.54x_5^2 + \frac{1}{-0.388 + 0.210/x_4^{0.488}}$$

All of the parameters including the flexibility parameter β_2 were significant at the 0.0001 level. Also, adding x_6 reveals that x_6 is not significant at the .05 level. Although the data contain no random error so that significance testing has no theoretical grounds here, the results of the significance tests do indicate the importance of the various effects relative to the ability of the overall model to fit the data. Note that the model contains only eight parameters but fits the 100 data points quite well as indicated by the corresponding cross validation plot given in Figure 6. The fact that the parametric nonlinear model does not fit the data as well as the

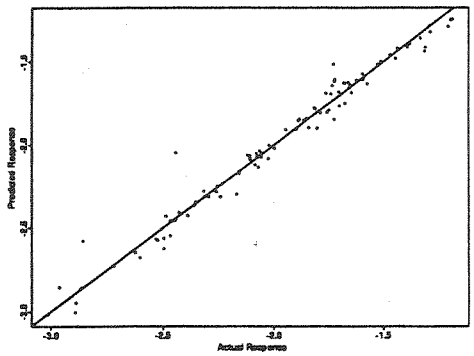

Figure 6: Cross Validation of the Parametric Nonlinear Model

nonparametric model is not surprising, however, since the parametric model is much simpler.

Some Comments on the Choice of Experimental Design

Originally, a 4^{6-2} fractional factorial design was considered for the solar collector computer experiment. A fractional factorial or even full factorial design would have had several drawbacks, however. First, if only a few covariates (factors) had an impact, the design effectively collapses into a smaller design with replications. But, replications in a computer experiment are non-informative because of the deterministic nature of the computer code and therefore would have been a waste of resources. Second, it could have been easy to miss an unknown effect by only experimenting at a few different points for each factor. For example, the exact nature of the nonlinear x_4 effect would have been difficult to identify with only four levels; in fact, the dramatic nonlinear behavior of x_4 surprised the engineers. Third, the decision of where to place the levels becomes much more crucial for the factorial design; lower dimensional projections of Latin hypercube design typically consist of n distinct and spread-out points so that their exact position is less important. Finally, a 4^{6-2} fractional factorial design would have required 256 runs. This compares to the 100-run Latin hypercube design that was used; even fewer runs might have been sufficient.

4 Discussion

The examples presented in nonlinear regression books typically deal with only a single covariate x, where the functional relationship between x and the response y is unknown. On the other hand, the method proposed in this paper can be applied to an arbitrarily large number of covariates.

There are certainly other ways to identify a parametric nonlinear model. For example, clever residual analyses in the hand of a skilled data analyst may well lead to the same results. For the solar collector experiment, an added variable plot for x_4 based on a linear regression model for the remaining covariates shows the effect of x_4 is nonlinear, albeit with considerable scatter as displayed in Figure 7. This success is not surprising since the assump-

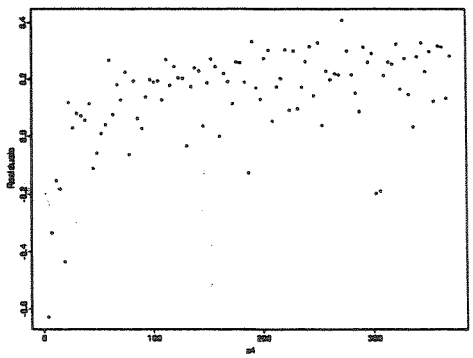

Figure 7: Added Variable Plot for x_4 in the Model $\text{logit}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5$

tion of a linear model for the remaining variables turns out to be a good approximation. If the true model had contained several strong nonlinearities, then added variable plots on their own would not have sufficed.

Elaborate residual analyses are often not done for three reasons: (1) They are hard to do, especially when the "true" model contains more than one nonlinear effect. (2) Data analysts, especially inexperienced ones, may not always know about them.

(3) They can take a lot of time to perform. The method presented here is easy and fairly automatic for detecting nonlinear effects. It is not a panacea for all "true" models, however. If the "true" model cannot be transformed to an additive model with few or no interaction effects, then identification of nonlinear relationships with several covariates will still be a challenge. For these cases, it is doubtful whether alternate methods will work either.

The effect plots play a key role in the proposed method and their resolution depends on the experimental design used. The Latin hypercube design is a desirable choice because the design points fill the experimental region well and produce high-resolution plots. Computer experiments typically use such space filling designs so the proposed method is particularly suited to computer experiments. While physical experiments typically collect much less data than computer experiments, in principle, the proposed method can be applied to physical experiments by adding a random error term to the model.

Acknowledgments

We thank S. Cao (now at the National Research Council of Canada) and T. Hollands, Solar Thermal Lab, Department of Mechanical Engineering, University of Waterloo for involving us in the solar collector design application. M. Hamada's research was supported by General Motors of Canada Limited, the Manufacturing Research Corporation of Ontario, and the Natural Sciences and Engineering Research Council of Canada. M. Schonlau and W. Welch's research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] Bates, D.M., and Watts, D.G. (1988), *Nonlinear Regression Analysis and its Applications*, New York: John Wiley & Sons.

[2] Cao, S. (1993), *Numerical Investigation on Unglazed Transpired Plate Solar Collector*, unpublished master's thesis, Department of Mechanical Engineering, University of Waterloo.

[3] Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), "Bayesian Prediction of Deterministic Functions, With Applications to the Design and Analysis of Computer Experiments," *JASA*, 86, 953-963.

[4] Golneshan, A. (1991), "Forced Convection Heat Transfer from a Transpired-Plate Having Slotted Perforations with Applications to Solar Air Heaters," unpublished Ph.D. proposal, Department of Mechanical Engineering, University of Waterloo.

[5] McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

[6] McKay, M.D., Beckman, R.J., and Conover, W.J. (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, 21, 239–245.

[7] Montgomery, D.C., and Peck, E.A. (1982) *Introduction to Linear Regression Analysis*, New York: John Wiley & Sons.

[8] Sacks, J., Schiller, S.B., and Welch, W.J. (1989), "Designs for Computer Experiments," *Technometrics*, 31, 41–47.

[9] Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409–435.

[10] Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P. , Mitchell, T.J., and Morris, M.D . (1992), "Screening, Predicting, and Computer Experiments," *Technometrics*, 34, 15–25.