A VISUAL INTERFACE FOR MODEL-FITTING

ANDREW J SMITH1, ZAHID MALIK2, JOHN NELDER3 AND ROBERT SPENCE2*

1Matra BAe Dynamics (UK), FPC 450, PO Box 5, Filton, Bristol BS34 7QW
2Department of Electrical and Electronic Engineering, Imperial College, London SW7 2BT, UK
3Department of Mathematics, Imperial College, London SW7 2BZ, UK

SUMMARY

We describe a tool which supports the activity of a human being in fitting a mathematical model to measured or simulated data. The tool offers two principal advantages; its use requires a minimum of statistical knowledge, and its visual and interactive nature ensures that its use is intuitive. The tool is novel in that, in the iterative and often exploratory construction of a model, it represents graphically the benefit of all possible single changes to the existing model; a single selection action suffices to add a new term to a model or to remove it. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: visualization; Response Surface Model; Generalized Linear Model; model fitting

1. INTRODUCTION

A common task for the scientist or engineer is that of fitting a mathematical model to a set of data. The data may originate from experimental measurements such as laboratory investigations and agricultural trials, or from simulations of an artefact carried out by computer; packages such as NASTRAN and SPICE are familiar simulation tools for the structural and electronic engineer, respectively. The mathematical relation eventually obtained will relate various input and output variables, typically referred to in the engineering domain as the parameters and performances of an artefact.

In this paper we focus on one particular form of mathematical model called a Response Surface Model (RSM). We take, for illustration, an example from the domain of engineering design, in which the parameters of an artefact must be chosen to satisfy some requirement on its performances. Such a model [1] is sufficiently simple to allow fast and inexpensive calculation, a feature which is especially valuable when simulation or measurement is expensive and/or time-consuming; it offers an attractive alternative to interpolation between known data values.

2. THE DEAP CYCLE

A common methodology available for model fitting is described by the ‘DEAP’ cycle [2] illustrated in Figure 1. In the Design phase, ‘Design of experiments’ techniques are used to plan which data points to collect in order to sample the model space adequately. The artefact is then measured or simulated in the Execution phase. The complementary tasks of model fitting and checking take place in the Analyse phase; typically, many iterations of exploratory fitting followed by checking take place in this phase, and it is this iterative behaviour which the proposed tool is intended to support. If the resulting model is adequate the process ends, otherwise the user Predicts which additional data points must be acquired, and the cycle of Figure 1 repeats.

3. MODELS

In the search for a RSM, a variety of Generalized Linear Models (GLMs) is available to the user to relate an output $y$ to a set of inputs $x_i$ [3]. The simplest is the polynomial model of the form

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2 + \cdots$$

where $a_i$ are the coefficients of the terms. However, a variety of alternatives to the polynomial model exists. A transformation (for example, inverse, hyperbolic, tan) may be applied to the input ($x_i$) and/or the performance variable in the model, and may well

*Correspondence to: R. Spence, Department of Electrical and Electronic Engineering, Imperial College, London SW7 2BT, UK. Email: r.spence@ic.ac.uk
Contract/grant sponsor: UK Engineering and Physical Sciences Research Council; Contract/grant number: GR/L31999

Copyright © 2001 John Wiley & Sons, Ltd.

Received 1 September 2000
Revised 30 November 2000
lead to a simpler and better predictive model. Inverse polynomial models have, for example, been shown to be particularly appropriate for modelling some electromagnetic devices [4,5]. As a further complication, certain coefficients in the polynomial may depend upon some categorical variables.

4. AVAILABLE FITTING TOOLS

Statistical packages such as Genstat [6] exist which allow a model form (e.g., polynomial) to be proposed and data to be provided. The best coefficient values (e.g., $a_1$, $a_2$, $a_{12}$) are calculated and a 'goodness of fit' measure provided, to indicate how closely the resulting RSM fits the available data. Genstat has very sophisticated and powerful algorithms for fitting models. This task can, however, be a time-consuming and difficult business for the non-statistician; in particular, it is difficult to 'see where to go next', and to be able to try different models rapidly. Genstat, and similar statistical packages, do not make it very easy to navigate 'model space'.

The disadvantage of a tool such as Genstat is that the user must understand the statistical theory and application very well. For this reason, use of the tool tends to be confined to professional statisticians and those few engineers who have an appropriate interest. A situation therefore tends to exist in which the statistician may well not possess useful domain knowledge (e.g., of electronic circuits) and the engineer may not have adequate statistical expertise. The advantage of the tool to be described is that it facilitates the acquisition of a useful model by a user who may have only minimal statistical expertise, but is equally valuable for the statistician.

5. THE TOOL

The interface of the Model Maker is shown in Figure 2. Each square represents a single term in the...
RSM being fitted, and is labelled accordingly; the terms are grouped according to their order. A term can exist in one of three possible states, and these are represented in the manner shown in Figure 3. A black circle inscribed within a square indicates that the corresponding term is already included within the model, whereas a white circle indicates that the term is not yet included. The size of these circles indicates the actual (black) or anticipated (white) value of the term’s inclusion within the model. A white background indicates that a term can be added to or removed from the model independently, whereas a grey background indicates that, for heuristic reasons (see below), the term’s inclusion requires other terms to be included at the same time. Thus, by employing the simple representations of state defined in Figure 3, the Model Maker supports iterative exploration from state (current model) to state (new model), by providing the contextual information concerning all possible changes ‘one step away’. The addition or removal of a term is accomplished by a single mouse-click. Note that there is no representation of a constant term—this is always included in the model.

The interface of the Model Maker also presents computed values of coefficients, standard errors, t-, F- and p-values, the goodness of fit (based on the deviance) and the current model. The latter two are also shown for the immediately previous model. Particularly valuable is a ‘history histogram’ (Figure 4) in which bar heights are related to a global performance parameter for the model, based on the deviance; the smaller the bar, the better the model.

5.1. Heuristics

A model fitting heuristic called Functional Marginality (FM) [3] is embedded within the Model Maker. The heuristic states that, for a particular term to be included in a model, all lower-order combinations of its component variables must also
be included. Thus, for $x_1^2x_2$ to be legal under FM, the terms $x_1$, $x_2$, $x_1^2$ and $x_1x_2$ would also need to be included. Omitting any of the latter ‘marginal terms’ is equivalent to making an assumption—concerning the underlying data—which is usually incorrect. The FM heuristic, which is checked automatically and requires no intervention from the user, underlies the use of the grey backgrounds in the Model Maker interface.

6. USE OF THE MODEL MAKER
The user of the Model Maker will typically begin by examining the first-order terms to identify those with the largest inscribed circles; these circles constitute ‘lookahead’ information helping to answer the question ‘Where should I go next?’ The size of a black circle is related to a term’s t-value, and that of a white circle to the relevant p-value. The selection of a term for inclusion in the model is achieved by a simple clicking action, as also is its removal. Model fitting is essentially an exploration from the current model to a new model, so the user will usually observe the history histogram to judge whether the most recent step taken was worthwhile; if not, it is possible to undo it, and therefore recover from a ‘blind alley’. Thus, during model fitting, the user is able to look ahead at all possible state changes and look back at long range contextual information—the ‘goodness of fit’ of all the models so far explored.

When the mouse cursor is placed over an element’s square, pink highlighting shows which (if any) other terms will be affected by clicking on that square. If the term is already included in the model, the highlighting indicates all higher-order terms which depend upon it according to the FM heuristic, and which would automatically be removed. Conversely, if the term is not currently included in the model, any missing marginal terms are highlighted, since they would automatically be added.

Two popular exploration strategies are supported by the Model Maker. One is a ‘top down’ strategy which begins with all terms present, whereupon the model is ‘pruned’. The other is a ‘bottom up’ strategy in which the model is built up from the constant term. Both strategies will often involve retreat to earlier models seen to be advantageous from examination of bar heights in the history histogram. Nevertheless, merely minimizing the bar height (the deviance) runs the danger of over-fitting—either modelling some of the noise in the data or creating an over-accurate model of part of the data space which does not hold for other regions. Despite this, the histogram is still a good and simple outline indication of model quality.

The ease with which sophisticated models can be explored aids problem holders with extensive domain knowledge (but not great statistical expertise). For example, different model classes can readily be chosen and fitted from the appropriate menu of transformations. These classes may be particularly suitable for the fitting, because they incorporate domain knowledge, physical limitations and other factors of which only the problem holder may be aware.

7. A STRUCTRUAL EXAMPLE
Figure 5 shows a structure, about 4 inches tall, whose function is to support the filament of an electric lamp [7]. The design of the structure involves the choice of values for the four-dimensional parameters $(x_1 \rightarrow a_1, x_2 \rightarrow a_2, x_3 \rightarrow d, x_4 \rightarrow r)$ to ensure that each of four stresses ($S_1$ to $S_4$) lies within acceptable bounds. To allow such a design to proceed, a RSM for each stress is required in terms of the four parameters.

Figure 6 shows one intermediate state in an exploration leading to a RSM for one of the stresses ($S_1$). It corresponds to a model of the form (appropriate coefficient terms would be calculated for each of the terms below so that $S_1 = a_0 + a_1x_1 + a_2x_2 + \cdots$)

$$S_1 \sim x_1 + x_2 + x_3 + x_4 + x_1^2$$

(1)

As can be seen from figure 6, there are other important terms which should be included. We can also see at once which are the most important, and hence which to investigate first. We can also display a plot of the fitted values against actual values, giving an easily understood graphical interpretation of how well our model fits the original data, and a plot of the residuals plotted against fitted values. These can give valuable information as to how well the model fits the data (principally how good our assumptions of the error distribution are), although some skill and training is required to interpret the residual plots correctly. The historical plot of the deviance show how quickly our model has improved as we included various terms. Finally, another window displays standard statistics associated with the current model.

The final model we settled on (Figure 7) was

$$S_1 \sim x_1 + x_2 + x_3 + x_4 + x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_3x_4 + x_1^2x_3 + x_1^2x_4 + x_2^2x_3 + x_1x_2^2 + x_2x_3^2 + x_1x_3^2 + x_2x_3$$

(2)

Figure 5. Philips lamp filament support system

Figure 6. Model Maker showing intermediate stage in the fitting of $S_1$
Some thought needs to be put into this model. We must remember that we only require a model which is ‘good enough’—all the models are merely approximations. We may therefore decide that some terms are not necessary, either from a physical/domain point of view or from the statistics. The histogram (and back buttons) are particularly useful here in showing that other models with fewer terms may be adequate. Ultimately, we seek a parsimonious model, and the Model Maker makes this task fast and straightforward.

8. COMMENT

Informal evaluation of the Model Maker has been encouraging, although no formal evaluation has yet taken place in view of the complexity involved in such an exercise. In practice, use or neglect of the tool will indicate its value, and for this reason it is made available on the Web (see Appendix A).

While described in the context of a polynomial model, the interface principle is equally relevant to any type of GLM. Various possible extensions to the tool are not difficult to identify; they include the support of richer GLMs, the incorporation of categorical variables, the pre-checking of data (for example, to eliminate variables which only contribute noise) and the detection of over-fitting [8].

ACKNOWLEDGEMENTS

We wish to thank Dr Lisa Tweedie for her initial design work for the Model Maker, and Simon Harding (IACR-Rothamsted, UK) for providing the Genstat libraries. The project was funded by the UK Engineering and Physical Sciences Research Council under grant GR/L31999.

APPENDIX A. HOW TO OBTAIN A COPY OF THE CODE

The most recent version of the code, together with a user-guide and sample data, can be found at:

http://infoeng.ee.ic.ac.uk/~smith/ajupload.html

For more information, please contact the first author (andrew-j.smith1@bae.co.uk).
REFERENCES


Authors’ biographies:

Andrew J. Smith obtained an MA (Hons) degree and a postgraduate Computer Science Diploma from Cambridge University. He spent 8 years at Imperial College, London, as a PhD student and then a Research Assistant. He currently works as a Principal Software Engineer at Matra-BAe Dynamics (UK) in Filton, Bristol. He is an Associate Member of the IEE (AMIIEE). His interests include HCI, pattern recognition, modelling and simulation.

Zahid Malik received his BSc and MSc from Imperial College in theoretical physics. After gaining a PhD in theoretical physics (Computer Simulation of Fundamental Quantum Processes…) from Portsmouth University he started working as a research associate in the Electrical Engineering Department of Imperial College and is now a Research Lecturer at the same institution. His research in the department has been focused on the statistical modelling, within a GLM framework, of computer simulations and the interactive visualization of the modelling process. He has several publications and conference presentations in this field.

John Nelder is visiting Professor in the department of Mathematics at Imperial College and a Fellow of the Royal Society. He was past president of the Royal Statistical Society and of the International Biometric Society. He was responsible for the development of the Genstat and GLIM packages. Professor Nelder was also responsible for the Generalized Linear Model approach and is well-known internationally for his contribution to the development of optimization algorithms (Nelder-Mead Simplex).

Robert Spence is Professor of Information Engineering at Imperial College, London and a Fellow of the Royal Academy of Engineering. His Research into electronic design and human-computer interaction is recorded, respectively, in about 80 and 60 publications. He drew upon both these interests as a co-architect of the MNNIE and CoCo computer-aided design systems. In particular, it is his aim to make statistical modelling techniques usable by a ‘problem holder’ within the overall engineering design process. Successful steps in this direction include the Attribute and Influence Explorers and the Prosection Matrix, all invented by Spence and his collaborators.