Journal of Statistical Planning and Inference 144 (2014) 55–62

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Robust static designs for approximately specified nonlinear regression models

Jamil Hasan Karami [a], Douglas P. Wiens [b,*]

[a] Department of Statistics Biostatistics & Informatics, University of Dhaka, Dhaka 1000, Bangladesh
[b] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1

ARTICLE INFO

Available online 13 September 2012

Keywords:
Approximate response
Bayesian optimality
Genetic algorithm
Minimax
Optimal design

ABSTRACT

We outline the construction of robust, static designs for nonlinear regression models. The designs are robust in that they afford protection from increases in the mean squared error resulting from misspecifications of the model fitted by the experimenter. This robustness is obtained through a combination of minimax and Bayesian procedures. We first maximize (over a neighborhood of the fitted response function) and then average (with respect to a prior on the parameters) the sum (over the design space) of the mean squared errors of the predictions. This average maximum loss is then minimized over the class of designs. Averaging with respect to a prior means that there is no remaining dependence on unknown parameters, thus allowing for static, rather than sequential, design construction. The minimization over the class of designs is carried out by implementing a genetic algorithm. Several examples are discussed.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Suppose that an investigator is planning an experimental study, the data from which will be analyzed by nonlinear regression. Thus, a specific parametric model will be fitted, and from this fit various inferences will be made. The design in such studies is typically chosen with the intention of minimizing some loss function such as the variance, or mean squared error, of the parameter estimates or the predicted values. A problem that arises immediately is that, by the nature of nonlinear models, these common measures depend on the unknown values of the parameters.

One way in which this problem might be handled is by constructing a 'locally optimal' design—one that is optimal only at a particular value $\boldsymbol{\theta}_0$ of the, often vector-valued, parameter. The designer hopes that the optimality will extend, at least approximately, to a neighborhood of $\boldsymbol{\theta}_0$. This parameter might arise from the experimenter's prior knowledge, or perhaps as an estimate from an earlier experiment.

A somewhat more robust approach allows for uncertainty about the parameter values, but not the form of the response function—one first maximizes the chosen loss function (or minimizes an efficiency measure such as the determinant of the information matrix) over a neighborhood of a local parameter $\boldsymbol{\theta}_0$ and then optimizes over the class of designs. See King and Wong (2000), Dette and Biedermann (2003), and Dette and Pepelyshev (2008) for examples of this approach.

With or without robustness considerations, a more practical method – if time allows – is to design sequentially. In this approach, evidently first proposed by Box and Hunter (1965a,b) parameter estimates are recalculated as the data accrue, and subsequent design points are computed in such a way as to optimize a performance measure, evaluated at the current

* Corresponding author.
E-mail addresses: karami@univdhaka.edu (J.H. Karami), doug.wiens@ualberta.ca (D.P. Wiens).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.09.001

estimates. Sinha and Wiens (2002) implement a robust variation of this method. They entertain a class of nonlinear models forming a neighborhood of that which the experimenter will fit, maximize the loss over this neighborhood, and then sequentially construct minimizing designs. We emphasize the 'full' nature of the neighborhoods over which they maximize – in contrast to the work referred to in the previous paragraph, the entire form of the response, rather than merely the parameter value, is allowed to vary. See also Sinha and Wiens (2003) for the asymptotic theory of this approach.

Yet another manner in which one can deal with this problem is through Bayesian optimality. Here the loss function is averaged, with respect to a 'prior' distribution on the parameters; this averaged value is then minimized. See Dette and Neugebauer (1997) for details. We note that it is not necessary to adopt any particular position on Bayesianism in order to take this approach; the introduction of a prior can be viewed merely as a convenient manner in which to eliminate the parameters from the loss function, thus allowing for static, i.e. nonsequential, design construction.

In this article we combine Bayesian optimality with robustness considerations as in Sinha and Wiens (2002), and go on to exhibit static designs with attractive robustness properties. Thus, in Section 2, we introduce a class of nonlinear models forming a neighborhood of that which the experimenter intends to fit. For each member of this class the 'true' parameter $\boldsymbol{\theta}_0$ is defined as that which minimizes an L_2-distance between the true regression response $E[Y|\mathbf{x}]$ and the fitted response $f(\mathbf{x}|\boldsymbol{\theta})$, as $\mathbf{x}$ ranges over a finite 'design space' $\mathcal{S}$. The loss is evaluated at $\boldsymbol{\theta}_0$ and maximized as

$$d(\mathbf{x}|\boldsymbol{\theta}_0) \stackrel{def}{=} E[Y|\mathbf{x}] - f(\mathbf{x}|\boldsymbol{\theta}_0), \tag{1}$$

ranges over an $L_2(\mathcal{S})$-neighborhood of *0*; the resulting maximum is then averaged with respect to a prior $p(\boldsymbol{\theta}_0)$.

In Section 3 we describe a genetic algorithm used to minimize the average maximized loss. Examples and further computational issues are presented and discussed in Section 4. Computing code, written in R, to duplicate these examples is available from us. Derivations are in the Appendix.

The designs constructed here are indexed by a parameter $v \in [0,1]$ representing the relative emphasis that the experimenter places on those errors, in the MSE of the fitted response, which are due to bias versus those due to random variation. When designing with $v = 0$ the experimenter is implicitly stating that he has complete faith in the fitted model $E[Y|\mathbf{x}] = f(\mathbf{x}|\boldsymbol{\theta}_0)$ and that only the parameter $\boldsymbol{\theta}_0$ is in doubt; the resulting designs are Bayesian optimal in the usual sense. It will be seen in the examples of Section 4 that as v increases the designs become more diffuse, hence more robust against errors in the specification of the functional form of the response. As $v \to 1$ the interpretation is that the experimenter has no faith in the accuracy of his specified model and so seeks a uniform exploration of the design space. There is of course a dependence of the design on the form of the prior, but this appears to be slight in comparison with the dependence on v.

Our designs are by construction 'exact', and so there is no difficulty in implementing them. As for their comparative performance—there is no shortage in the literature of simulation studies comparing robust with nonrobust designs; these all tend to give the same, unsurprising message which need not be dwelt upon here. The robust designs give more protection, against model misspecifications, than do the nonrobust designs tailored for efficiency at one particular model, and the latter tend to be more efficient when the fitted model is indeed exactly correct. See for example the case studies in Fang and Wiens (2000) and in Li and Wiens (2011).

2. Approximately specified nonlinear models

In this section we describe our scenario for design in an approximately specified nonlinear regression framework. A 'design' is a specification $(n_1/n, \ldots, n_N/n)$, for integers $\{n_i\}$ summing to n. In implementing the design the experimenter makes n_i observations $\{Y_{ij}\}_{j=1}^{n_i}$ at 'location' $\mathbf{x}_i$—a q-dimensional vector of covariates chosen from a design space $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^{N}$. We note that N, while finite, may be arbitrarily large, and that $n_i \geq 0$—there is no requirement that observations be made at every design point. After sampling, a known – with some reservations – regression response function $f(\mathbf{x}|\cdot)$ will be fitted, viz. $\hat{Y}(\mathbf{x}) = f(\mathbf{x}|\hat{\boldsymbol{\theta}})$. The approximate nature of the model – that $E[Y|\mathbf{x}] \approx f(\mathbf{x}|\boldsymbol{\theta}_0)$ for some p-dimensional $\boldsymbol{\theta}_0$ – is characterized by first defining

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} (E[Y|\mathbf{x}_i] - f(\mathbf{x}_i|\boldsymbol{\theta}))^2. \tag{2}$$

Then with $d(\mathbf{x}|\boldsymbol{\theta}_0)$ given by (1) the probability model is

$$Y(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}_0) + d(\mathbf{x}|\boldsymbol{\theta}_0) + \varepsilon,$$

with random errors ε. We assume independent, homoscedastic errors, with variance σ^2.

Assume that $f(\mathbf{x}|\cdot)$ is differentiable with respect to $\boldsymbol{\theta}$, define $\mathbf{Z}(\boldsymbol{\theta})$ to be the $N \times p$ matrix with i^{th} row

$$\mathbf{z}^T(\mathbf{x}_i|\boldsymbol{\theta}) = \frac{\partial f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

and assume that $\mathbf{Z}(\boldsymbol{\theta}_0)$ has full column rank. A consequence of (2) is that, if $\mathbf{d}(\boldsymbol{\theta}) = (d(\mathbf{x}_1|\boldsymbol{\theta}), \ldots, d(\mathbf{x}_N|\boldsymbol{\theta}))^T$, then

$$\mathbf{Z}^T(\boldsymbol{\theta}_0)\mathbf{d}(\boldsymbol{\theta}_0) = \mathbf{0}_{p\times 1}. \tag{3}$$

Download English Version:

https://daneshyari.com/en/article/1148605

Download Persian Version:

https://daneshyari.com/article/1148605

Daneshyari.com