Applied Mathematics Letters 43 (2015) 10–18

Contents lists available at ScienceDirect

Applied Mathematics Letters

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

Model comparison tests to determine data information content

H.T. Banks [a,*], J.E. Banks [b], Kathryn Link [a], J.A. Rosenheim [c,d], Chelsea Ross [a], K.A. Tillman [a]

[a] Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8212, USA

[b] Division of Sciences & Mathematics, School of Interdisciplinary Arts & Sciences, University of Washington, Tacoma, Tacoma, WA 98402, USA

[c] Department of Entomology and Nematology, University of California, Davis, Davis, CA 95616, USA

[d] Center for Population Biology, University of California, Davis, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 4 November 2014
Accepted 4 November 2014
Available online 28 November 2014

Keywords:
Ordinary least squares
Model comparison in inverse problems
Information content

ABSTRACT

In the context of inverse or parameter estimation problems we demonstrate the use of statistically based model comparison tests in several examples of practical interest. In these examples we are interested in questions related to information content of a particular given data set and whether the data will support a more complicated model to describe it. In the first example we compare fits for several different models to describe simple decay in a size histogram for aggregates in amyloid fibril formation. In a second example we investigate whether the information content in data sets for the pest *Lygus hesperus* in cotton fields as it is currently collected is sufficient to support a model in which one distinguishes between nymphs and adults. Finally in a third example with data for patients having undergone an organ transplant, we question whether the data content is sufficient to estimate more than 5 of the fundamental parameters in a particular dynamic model.

1. Introduction

Uncertainty quantification in the context of estimation of parameters has become a focus of increased attention in recent years. As mathematical models become more complex with multiple states and many parameters to be estimated using experimental data, there is a need for critical analytical tools in model validation related to the reliability of parameter estimates obtained in model fitting. Methodology is desirable to distinguish between lack of identifiability in a model (often formulated in a generalized algebraic context) vs. local insensitivity with respect to changes in particular parameters vs. lack of information content in a given data set. A recent concrete example involves previous HIV models [1,2] with 15 or more parameters to be estimated. In [3], using recently developed parameter selectivity tools [4] based on parameter sensitivity based scores, the authors showed that many of the parameters could not be estimated with any degree of reliability. Moreover, it was found that quantifiable uncertainty varies among patients depending upon the number of treatment interruptions (perturbations of therapy). This leads to a *fundamental question* of how much information with respect to model validation can be expected in a given data set or collection of data sets. In this note, we consider one tool that may be used in attempts to answer this question.

* Corresponding author.
E-mail address: htbanks@eos.ncsu.edu (H.T. Banks).

http://dx.doi.org/10.1016/j.aml.2014.11.002
0893-9659/

Here we demonstrate the use of *statistically based model comparison tests* in several examples of practical interest. In these examples we are interested in questions related to information content of a particular given data set and whether the data will support a more detailed or sophisticated model to describe it. In the first example we compare fits for several different models to describe simple decay in a size histogram for aggregates in amyloid fibril formation. In a second example we investigate whether the information content in data sets for the pest *Lygus hesperus* in cotton fields as it is currently collected is sufficient to support a model in which one distinguishes between nymphs and adults. Finally in a third example with data for patients having undergone an organ transplant we question whether the data content is sufficient to estimate more than 5 of the fundamental parameters in a specific dynamic model. In the next section we recall the fundamental tests to be employed here.

2. Summary of ANOVA type statistical comparison tests

In general, assume that we have an inverse problem for the model observations $f(t, q)$ and are given n observations. We define

$$J_n(q) = J_n(\mathbf{Y}, q) = \frac{1}{n}\sum_{j=1}^{n}[Y_j - f(t_j, q)]^2 \tag{1}$$

where our statistical model has the form

$$Y_j = f(t_j, q_0) + \mathcal{E}_j, \quad j = 1, \ldots, n.$$

Here, q_0 is the "true" value of q which we assume to exist. We use $\mathcal{Q}$ to represent the set of all the admissible parameters q.

We make the standard statistical assumptions [5–7]:

- (A1) The random variables $\{\mathcal{E}_j\}_{j=1}^{\infty}$ are independent and identically distributed with $\mathbb{E}(\mathcal{E}_j) = 0$ and $Var(\mathcal{E}_j) = \sigma^2$.
- (A2) $\mathcal{Q}$ is a compact subset of Euclidean space of R^p and $f(t, q)$ is continuous on $[0, T] \times \mathcal{Q}$.
- (A3) Observations are at $\{t_j\}_{j=1}^{n}$ in $[0, T]$. For some finite measure μ on $[0, T]$,

$$\frac{1}{n}\sum_{j=1}^{n} h(t_j) \longrightarrow \int_0^T h(t)d\mu(t)$$

  as $n \to \infty$, for all continuous functions h.
- (A4) $J_0(q) = \int_0^T (f(t, q_0) - f(t, q))^2 d\mu(t) = \sigma^2$ has a unique minimizer in $\mathcal{Q}$ at q_0.

Let $q^n = q^n_{OLS}(\mathbf{Y})$ be the OLS estimator for J_n with corresponding estimate

$$\hat{q}^n = q^n_{OLS}(\{y_j\})$$

for a realization $\mathbf{y} = \{y_j\}$. That is,

$$q^n(\mathbf{Y}) = \arg\min_{q\in\mathcal{Q}} J_n(\mathbf{Y}, q)$$

and

$$\hat{q}^n = \arg\min_{q\in\mathcal{Q}} J_n(\mathbf{y}, q).$$

One can then establish a series of useful results (see [5,6] for detailed proofs; see also [8]).

Theorem 2.1. *Under* (A1)–(A4), $q^n = q^n_{OLS}(\mathbf{Y}) \longrightarrow q_0$ *as* $n \to \infty$ *with probability* 1.

We will need further assumptions to precede (these will be denoted by (A7)–(A11) to facilitate reference to [5,6]). These include:

- (A7) $\mathcal{Q}$ is finite dimensional in R^p and $q_0 \in \text{int } \mathcal{Q}$.
- (A8) $f : \mathcal{Q} \to C[0, T]$ is a C^2 function.
- (A10) $\mathcal{J} = \frac{\partial^2 J_0}{\partial q^2}(q_0)$ is positive definite.
- (A11) $\mathcal{Q}_H = \{q \in \mathcal{Q} | Hq = c\}$ where H is an $r \times p$ matrix of full rank, and c is a known constant.

In many instances, including the motivating examples discussed here, one is interested in using data to question whether the "true" parameter q_0 can be found in a subset $\mathcal{Q}_H \subset \mathcal{Q}$ which we assume for discussions here is defined by the constraints of assumption (A11). Thus, we want to test the *null hypothesis* H_0: $q_0 \in \mathcal{Q}_H$.

Define then

$$q^n_H(\mathbf{Y}) = \arg\min_{q\in\mathcal{Q}_H} J_n(\mathbf{Y}, q)$$

and

$$\hat{q}^n_H = \arg\min_{q\in\mathcal{Q}_H} J_n(\mathbf{y}, q)$$