Computational Statistics and Data Analysis 113 (2017) 226–238

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Model selection via Bayesian information capacity designs for generalised linear models

David C. Woods [a,*], James M. McGree [b], Susan M. Lewis [a]

[a] University of Southampton, UK

[b] Queensland University of Technology, Australia

ARTICLE INFO

Article history:
Received 29 January 2016
Received in revised form 22 October 2016
Accepted 26 October 2016
Available online 2 November 2016

Keywords:
Bayesian *D*-optimality
Factorial experiments
Generalised information criterion
Screening

ABSTRACT

The first investigation is made of designs for screening experiments where the response variable is approximated by a generalised linear model. A Bayesian information capacity criterion is defined for the selection of designs that are robust to the form of the linear predictor. For binomial data and logistic regression, the effectiveness of these designs for screening is assessed through simulation studies using all-subsets regression and model selection via maximum penalised likelihood and a generalised information criterion. For Poisson data and log-linear regression, similar assessments are made using maximum likelihood and the Akaike information criterion for minimally-supported designs that are constructed analytically. The results show that effective screening, that is, high power with moderate type I error rate and false discovery rate, can be achieved through suitable choices for the number of design support points and experiment size. Logistic regression is shown to present a more challenging problem than log-linear regression. Some areas for future work are also indicated.

1. Introduction

An important problem in scientific discovery is to find those variables (or factors) that have a substantive influence on an observed response through experiments on a possibly large set of potentially important variables. There has been much research into such variable screening, or model selection, focussed on the design and analysis of experiments in which the response variable is adequately approximated by a linear model (see Draguljić et al., 2014 and Woods and Lewis, 2016, and references therein). Such experiments are used increasingly in scientific research and product development, for example, in the pharmaceutical and chemical industries.

In many practical applications, for example when binary or count data are observed, a generalised linear model (GLM; McCullagh and Nelder, 1989) may be needed to describe a response. Previous research on designs for model selection for GLMs has focussed on experiments involving only a few variables through pairwise comparisons of a small number of models (see, for example, López-Fidalgo et al., 2007 and Waterhouse et al., 2008). Hence, such methods are not applicable to, or easily generalisable for, the screening problem. In the literature, the majority of multi-variable experimentation with GLMs has employed (fractional) factorial designs, including examples on solder-joint defects (Hamada and Nelder, 1997), windshield moulding, non-conforming tiles and semi-conductor defects (see Lewis et al., 2001). Although such designs are effective for both model selection and estimation for normal-theory linear models, they have been shown to be inefficient for experiments that provide non-normal data (Woods et al., 2006).

* Correspondence to: Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK.
E-mail address: D.Woods@southampton.ac.uk (D.C. Woods).

http://dx.doi.org/10.1016/j.csda.2016.10.025

In common with other non-linear models, for the GLMs considered in this paper the performance of a design depends on the unknown values of the parameters in the model. One approach to overcoming this problem is to assume a particular value for each parameter and hence obtain a "locally optimal" design; that is, a design that is optimal under a given criterion provided the assigned parameter values are correct. We adopt the alternative approach of making the less stringent assumption of a prior distribution for each model parameter from which we obtain a "pseudo-Bayesian" design (Atkinson and Woods, 2015).

In this paper, we investigate variable screening for GLMs with q independent variables, labelled $x_1, \ldots, x_q$. In the jth run ($j = 1, \ldots, N$) of the experiment, a treatment or combination of variable values $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{qj})^{\mathrm{T}}$ is applied to an experimental unit and a univariate response, y_j, is observed. We assume that $|x_{ij}| \leq 1$ for $i = 1, \ldots, q$; $j = 1, \ldots, N$.

The aim of the experiment is to identify those *active* variables having a substantial effect on the response variable and to estimate efficiently a GLM involving those variables alone. For $j = 1, \ldots, N$, the y_j have independent exponential family distributions with expectation μ_j related to a linear predictor $\eta_j = f(\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{\beta}$ via a link function, $g(\mu_j) = \eta_j$. The vectors $f(\boldsymbol{x})$ and $\boldsymbol{\beta}$ are $p \times 1$ vectors of known functions of $\boldsymbol{x}$ and unknown model parameters, respectively. We also assume that the experimental units are exchangeable, in the sense that the distribution of the response to a treatment does not depend on the unit to which the treatment is applied.

For canonical link functions, the log-likelihood may be written as

$$l(\boldsymbol{\beta};\, \boldsymbol{y}) = \sum_{j=1}^{N} \left[y_j \eta_j - b(\eta_j) + c(y_j) \right], \tag{1}$$

where $b(\cdot)$ and $c(\cdot)$ are known functions of the linear predictor and response, respectively. For the binomial distribution and the logistic link, $b(\eta_j) = -n_j \log(1 + e^{\eta_j})$ and $c(y_j) = \log(n_j!/[y_j!(n_j - y_j)!])$, with n_j the number of Bernoulli trials made at the jth run. For the Poisson distribution and the log link, $b(\eta_j) = e^{\eta_j}$ and $c(y_j) = -\log(y_j!)$.

Maximum likelihood estimators (MLEs) $\hat{\boldsymbol{\beta}}$ can be found via (numerical) maximisation of (1). For small data sets, however, the MLEs may have considerable bias. For sparse data, such as binomial data with small numbers, n_j, of trials for each run, one or more maximum likelihood estimates may be infinite, for example, as the result of separation of the responses into zeros and ones via a hyperplane in the linear predictor (Silvapulle, 1981). To remove this bias and guarantee the existence of estimates for GLMs with a canonical link function, Firth (1993) defined penalised maximum likelihood estimators $\tilde{\boldsymbol{\beta}}$ as maximisers of

$$l^{\star}(\boldsymbol{\beta};\, \boldsymbol{y}) = l(\boldsymbol{\beta};\, \boldsymbol{y}) + \frac{1}{2} \log \det \left\{ X^{\mathrm{T}} W X \right\}, \tag{2}$$

where X is the $N \times p$ model matrix with jth row $f(\boldsymbol{x}_j)^{\mathrm{T}}$ and $W = \mathrm{diag}\{\mathrm{var}(y_j)\}$ (see also Kosmidis and Firth, 2009). This estimation procedure is equivalent to finding the posterior mode of $\boldsymbol{\beta}$ assuming the Jeffreys prior distribution.

The information matrix $X^{\mathrm{T}}WX$, which is the asymptotic inverse variance–covariance matrix for both $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, is used to define the D-optimality criterion. This criterion specifies selection of a design that maximises the objective function

$$\phi_D(\xi) = \frac{1}{p} \log \det \left\{ X^{\mathrm{T}} W X \right\}, \tag{3}$$

where

$$\xi = \begin{Bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \\ \omega_1 & \cdots & \omega_n \end{Bmatrix}, \tag{4}$$

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the distinct treatments in the design (assumed, without loss of generality, to be applied to the first n runs of the experiment), $\omega_k > 0 \in \mathbb{N}$, and $\sum_{k=1}^{n} \omega_k = N$, the total number of runs. For the GLMs considered in this paper, (3) depends on $\boldsymbol{\beta}$ through the matrix W and hence selection of a D-optimal design requires knowledge of the values of these parameters. Thus a locally optimal design is obtained.

The relative performance of two designs, ξ_1 and ξ_2, under D-optimality may be assessed using relative D-efficiency, defined as

$$\mathrm{DEff}(\xi_1, \xi_2) = \exp\left\{ \phi_D(\xi_1) - \phi_D(\xi_2) \right\}, \tag{5}$$

where $0 \leq \mathrm{DEff}(\xi_1, \xi_2)$. If ξ_2 is a D-optimal design that maximises (3), then (5) provides an absolute measure of the performance of design ξ_1.

In this paper, we address the screening problem of model selection and estimation of parameters in the selected model. We define, in Section 2, a Bayesian information capacity criterion that generalises D-optimality to provide model-robust designs for GLMs. We also present and discuss a model selection strategy that uses all-subsets regression and suitable penalties for model complexity. Sections 3 and 4 describe simulation studies of logistic and log-linear regression modelling, respectively, which demonstrate and assess the effectiveness of the methods. In Section 5, we present some avenues for future work to further develop methodology for screening experiments with non-normal data.