Linear Algebra and its Applications 536 (2018) 152–173

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Projections of a general binary model on a logistic regression

Mariusz Kubkowski [a], Jan Mielniczuk [a,b,*]

[a] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland*
[b] *Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland*

ARTICLE INFO

Article history:
Received 22 March 2017
Accepted 14 September 2017
Available online 19 September 2017
Submitted by W.B. Wu

MSC:
62J12
62F99

Keywords:
First canonical vector
General binary model
Additive binary model
Logistic regression
Misspecification
Kullback–Leibler divergence
Stein lemma

ABSTRACT

We consider a general binary model for which conditional probability of success given vector of predictors $\mathbf{X}$ equals $q(\beta_1^T\mathbf{X}, \ldots, \beta_k^T\mathbf{X})$ and a family of possibly misspecified logistic regressions fitted to it. In the case when $\mathbf{X}$ satisfies linearity condition we show that their algebraic structure is uniquely determined and that the vector β^* corresponding to Kullback–Leibler projection on this family is a linear combination of $\beta_1, \ldots, \beta_k$. This generalizes the known result proved by P. Ruud for $k = 1$ which says that the true and projected vectors are collinear. It also follows that the projected vector has the same direction as the first canonical vector which justifies frequent observations that logistic fit yields well performing classifiers even if misspecification is expected. In the special case of additive binary model with multivariate normal predictors and when response function q is a convex combination of univariate responses we show that the variance of $\beta^{*T}\mathbf{X}$ is not larger than the maximal variance of the projected linear combinations for the corresponding univariate problems. In the case of balanced additive logistic model it follows that the contribution of β_i to β^* is bounded

[*] Corresponding author at: Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland.
E-mail addresses: M.Kubkowski@mini.pw.edu.pl (M. Kubkowski), jan.mielniczuk@ipipan.waw.pl (J. Mielniczuk).

http://dx.doi.org/10.1016/j.laa.2017.09.013

by the corresponding coefficient in the convex representation of response function q.

1. Introduction

We consider a general binary model for which probability of a positive outcome $Y = 1$ given vector of random predictors $\mathbf{X} = (X_1, \ldots, X_p)^T$ is semi-parametrically modelled as

$$\mathbf{P}(Y = 1|\mathbf{X}) = q(\beta_1^T\mathbf{X}, \ldots, \beta_k^T\mathbf{X}), \tag{1}$$

where q is an unknown response function, $k \leq p$ and $\beta_1, \ldots, \beta_k \in \mathbf{R}^p$ are unknown column vectors of parameters. This is the parsimonious model yielding a flexible approach to binary dependence which is used frequently for dimension reduction and which encompasses e.g. the projection pursuit regression model (cf. e.g. [1]). It is known that equation (1) is equivalent to an apparently more general equality that $Y = h(\beta_1^T\mathbf{X}, \ldots, \beta_k^T\mathbf{X}, \varepsilon)$, where ε is a random variable independent of $\mathbf{X}$ (cf. [2], Lemma 1). In the following we will use vectorized version of the response function and write $q(\mathbf{B}^T\mathbf{X}) = q(\beta_1^T\mathbf{X}, \ldots, \beta_k^T\mathbf{X})$, putting $\mathbf{B} = [\beta_1, \ldots, \beta_k] \in \mathbf{R}^{p\times k}$. Note that we follow the usual convention (cf. [1]) of not explicitly defining the intercepts in (1). They may be included in the model by a suitable modification of function q.

There are several approaches to estimate sufficient dimension reduction directions $\beta_1, \ldots, \beta_k$, the most popular being Sliced Inverse Regression (SIR) method developed by [1] and [3], see [4] for recent developments. Here we consider inference issues arising when model (1) is misspecified as logistic regression model. Namely, to the data pertaining to (1) we fit the logistic regression model i.e. we postulate that the posterior probability that $Y = 1$ given $\mathbf{X} = x$ is of the form

$$q_L(\gamma_0 + x^T\gamma) = \exp(\gamma_0 + x^T\gamma)/[1 + \exp(\gamma_0 + x^T\gamma)], \tag{2}$$

where $\gamma_0 \in \mathbf{R}, \gamma \in \mathbf{R}^p$ are parameters. Our main interest here is the situation when the logistic model (2) is misspecified i.e. when $k \neq 1$ or when $k = 1$ but $q(s) = q_L(as + b)$ for all $s \in \mathbf{R}$ does not hold for any $a, b \in \mathbf{R}$. Obviously, when $k = 1$ and $q \equiv q_L$ we consider fitting of the logistic regression to a correctly specified conditional distribution. The problem has been studied for $k = 1$ with important contributions by [5], [6] and [7] among others, for a recent contribution and more references see [8]. The logistic model (2) is an ubiquitous modelling tool, however frequently it is applied without convincing evidence that it is adequate. We thus believe that misspecification case is common and its consequences are worth studying.