To appear in *Statistics: A Journal of Theoretical and Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–25

# *Active sets of predictors for misspecified logistic regression*

M. Kubkowski[a] and J. Mielniczuk[a,b]*

[a] *Faculty of Mathematics and Information Science, Warsaw University of Technology;*
[b] *Institute of Computer Science Polish Academy of Sciences*

(*v. 1.0 April 2016*)

We consider the case of a binary regression model to which a logistic regression is erroneously fitted. We investigate the interplay between the support $t$ of the true parameter and the support $t^*$ of its Kullback-Leibler projection on the logistic model. The objectives of the paper are to prove a new positive result specifying conditions under which $t^*$ is subset of $t$ and to show that any interplay between those two sets is possible in general. Moreover, we treat in detail the important special case when the true parameter and its projection are proportional and show among others how the projection on the full set of predictors and its subsets relate. The situation of simultaneous fit of the logistic and the linear model is also considered and it is shown that for the normal predictors direction of the fitted logistic regression parameter can be recovered from the corresponding linear fit.

**Keywords: misspecification, binary and logistic model, active set of predictors, Kullback-Leibler projection, response function**

*AMS Subject Classification*: Primary 62J02, Secondary 62J12

## 1. Introduction

We consider a general binary regression model such that a conditional distribution of $Y$ given $X$ for a random vector $(X, Y) \in R^{p+1} \times \{0, 1\}$ is given by

$$P(Y = 1|X = x) = q(x^T\beta), \tag{1}$$

where vector $X = (1, X_1, \ldots, X_p)^T$ are predictors, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is an unknown vector of parameters and $q : R \to (0, 1)$ is a certain unknown response function. Note that $X_0 \equiv 1$ corresponds to a constant term in regression equation (1). Moreover, we assume that predictors $X_1, \ldots, X_p$ are random. Observe that if $q$ is a cumulative distribution i.e. nondecreasing right-continuous function with limits 0 and 1 at respective infinities then the binary model can be written in the form

$$Y = I\{X^T\beta \geq \varepsilon\} =: g(X^T\beta, \varepsilon), \tag{2}$$

where $\varepsilon$ is r.v. which is independent of $X$ and has cumulative distribution function $q$. For a general $g$ such models are called single index models in the literature.
To the data pertaining to (1) we fit the logistic regression model i.e. it is assumed that

*Corresponding author. Email: miel@ipipan.waw.pl

the posterior probability that $Y = 1$ given $X = x$ is of the form

$$q_L(x^T b) = \exp(x^T b)/[1 + \exp(x^T b)], \tag{3}$$

where $\gamma \in R^{p+1}$ is a parameter. Obviously, when $q \equiv q_L$ we consider fitting of the logistic regression to a correctly specified conditional distribution. Let

$$t = \{0\} \cup \{1 \leq k \leq p : \beta_k \neq 0\}$$

be the set of indices of all active variables augmented by an index of the intercept denoted by 0. An important statistical problem is selection, which amounts to choosing data-based selector $\hat{t}$ such that it approximates $t$ in a certain specified sense. This is an area of intensive research, especially when $p$ is large, possibly larger than a sample size, and regression is sparse in the sense that number of relevant variables is much smaller than $p$. Recent representative examples of solutions are given in papers [5], [9], [15] and monographs [3], [13],[23] contain the overview of the field.
Consider the approach to select $t$ based on Maximum Likelihood (ML) estimation and define log-likelihood under (3)

$$l(b, X, Y) = Y \log q_L(X^T b) + (1 - Y) \log(1 - q_L(X^T b)) = Y X^T b - \log(1 + \exp(X^T b) \tag{4}$$

Let $R(b) = -E_{(X,Y)} l(b, X, Y)$ be the corresponding risk function. An object of main interest here is the minimizer of the risk

$$\beta^* = \text{argmin}_{b \in R^p} R(b), \tag{5}$$

which in view of (4) can be equivalently written as

$$\beta^* = \text{argmin}_{b \in R^p} E\Delta_X(q(X^T\beta), q_L(X^T b)), \tag{6}$$

where

$$\Delta_X(q(X^T\beta), q_L(X^T b)) = \\ q(X^T\beta) \log\left(\frac{q(X^T\beta)}{q_L(X^T b)}\right) + (1 - q(X^T\beta)) \log\left(\frac{1 - q(X^T\beta)}{1 - q_L(X^T b)}\right). \tag{7}$$

Note that the quantity in (7) is Kullback-Leibler (KL) distance between two Bernoulli binary variables with success probabilities equal $q(X^T\beta)$ and $q_L(X^T b)$, respectively. Thus $\beta^*$ is the minimizer of the averaged KL distance. When $q \equiv q_L$, in view of information inequality we have that $\beta^* = \beta$. We call $\beta^*$ Kullback-Leibler (KL) projection of $\beta$ and argue below that its support

$$t^* = \{i : \beta_i^* \neq 0\} \cup \{0\}$$

plays an important role in model selection for misspecified binary models. Note that $t^*$ can be interpreted as a set of indices of an active set of variables for the best fitted model augmented by 0. We review first known results concerning KL projection $t^*$ and its relevance in estimation and selection.
Assume momentarily that $p < n$ and let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample

pertaining to $P_{(X,Y)}$. Define ML estimators $\hat{\beta}_n$ as minimizers of an empirical risk $R_n(b) = -n^{-1}\sum_{i=1}^{n} l(b, X_i, Y_i)$, namely

$$\hat{\beta}_n = \text{argmin}_{b\in R^p} R_n(b). \tag{8}$$

Intuitively, as $\hat{\beta}_n$ and $\beta^*$ are minimizers of the empirical and the population risk, respectively, and $R_n(\cdot)$ is usually close to $R(\cdot)$, estimator $\hat{\beta}_n$ should be close to $\beta^*$. This is indeed the case for fixed $p$ and under some conditions due to concavity of the risk function, see e.g. [14]. Then $\hat{\beta}_n$ is consistent estimator of $\beta^*$ (see [8]). Thus the question of how support $t^*$ of $\beta^*$ relates to support of $t$ becomes important since selection procedures based on $\hat{\beta}_n$ or other estimators will approximate $t^*$ and not $t$. This was recognised long ago and some results on interplay between $t$ and $t^*$ have been established. In particular, important result of Ruud ([21], see also [16]) states that if regressions of $X$ given $X^T\beta$ are linear functions of the condition then $\tilde{\beta}^* = \eta\tilde{\beta}$ , where $\tilde{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ and $\tilde{\beta} = (\beta_1, \ldots, \beta_p)^T$. Thus either $t^* = t$ or $t^* = \{0\}$ if $\eta = 0$. Note that Ruud's result implies that if $\eta \neq 0$ then the directions of $\tilde{\beta}^*$ and $\tilde{\beta}$ are the same which is all we can hope for under misspecification, as a change of a scale of the unknown response results in a change of $\tilde{\beta}$ by a multiplicative constant.

In the follow-up paper [12] it is shown that when the number of predictors grows to infinity, approximate linearity of projection holds for large set of $\beta$ on the unit sphere, suggesting that the Ruud's result also approximately holds for such $\beta$. This, however, does not settle the question what is the interplay between $\beta$ and $\beta^*$ for a specific $\beta$.

It is shown in [19] that monotonicity of the response function is an important additional condition in the sense that under some mild additional assumptions $t^*$ is non-empty provided that $t$ is. Thus under Ruud's condition it means that $\eta \neq 0$ and whence $t = t^*$. In [19] two-step selection method $\hat{t}^*$ based on Generalized Information Criterion is also constructed which yields $P(\hat{t}^* \neq t^*) \to 0$ for fixed $p$ when sample size grows. We also note that (5) can be equivalently viewed as the problem of finding the minimal risk classifier for a logistic loss. We refer to [24] for bounds on pertaining regrets for general losses.

The important case of misspecification is omission of several variables in logistic regression model in which case the response function is an integral of the logistic response with respect to conditional density of omitted variables given the retained ones. It is easy to see that when the omitted variables are independent from the rest, the resulting model satisfies (1). This special case was treated e.g. in [11] and [20]. In particular it is proved in [11] that omission of important predictors from logistic model leads to underestimation of the effect of relevant binary predictor even if all observations are randomized with respect to the omitted covariates. The extreme case of such situation is filtering when individual predictors are fitted to the response and all others are omitted. There, omission is purposefully done to screen out inactive variables and by doing this to reduce the pool of potentially important predictors. In such a case we want to know whether the variables in the full model will be relevant when fitted as separate predictors to the response, as this ensures that we will not overlook any important ones. For a representative example of such methods see e.g. [10]. Other examples of naturally occurring misspecified binary models are discussed in [6] where the case of the logistic response misclassified with probability $\tau$ is considered. This leads to the response function $q(s) = (1-\tau)q_L(s)+\tau(1-q_L(s))$. Note that in cases when the response function is a convex combination of two or more logistic response functions the effects of misspecification could be mitigated by fitting such model and estimating corresponding parameters. In [4] fitting of misspecified linear model to a nonlinear one is considered. Although the

focus of the paper is on inference, some interesting results concerning $t$ and $t^*$ are proved there. In particular, it is shown in Proposition 3 of [4] that $t^* \subseteq t$ for Gaussian predictors, which means that an active variable in a misspecified linear model is relevant in explaining relationship between the predictors and the response.

Our main objective here is to study the interplay between $t$ and $t^*$ in the situation when the logistic model (3) is misspecified i.e. where for any $a, b \in R$ there exists $s \in R$ such that $q(s) \neq q_L(as + b)$. The problem of relating $t$ and $t^*$ for fixed number of predictors is scarcely studied partly due to technical difficulty of solving normal equation (12) below. Apart from Ruud's result and some partial results in [19] not much is known about it, despite its importance for vital statistical problems such as selection and filtering. In particular, active variables omitted from the fit may be replaced with inactive ones which are correlated with them and included in the modelling. Quantification of this phenomenon deserves more extensive research.

In Section 3 of the paper we prove some new results on the interplay between the active set of variables of the true model and the active set of its KL projection or, equivalently on active sets of indices $t$ and $t^*$. In particular our Theorem 3.1 implies that when the block of spurious variables $X_2$ is omitted from predictors $X = (X_1^T, X_2^T)^T$ then $t^*$ does not change provided that $X_2 - AX_1$ is independent of $X_1$ for some linear transformation $A$. This in particular implies that when $X_2$ is exactly the set of active predictors then $t^* \subseteq t$. We also show in Theorem 3.7 that $\beta^*$ enjoys certain continuity property and prove that for any response function $q$ we can construct distribution of predictors such that the set of active predictors and its KL projection have only the intercept in common. Surprisingly, such situation occurs also for the case when $q$ is monotone disproving the conjecture that in this case $t^* \subseteq t$.

We also consider in more detail the important situation when predictors satisfy a certain linear regression condition (cf. 20)), which we call Rudd's condition, and give expression for proportionality constant $\eta$ in Proposition 3.8. We prove that a simple condition $\mathrm{Cov}(Y, X^T\beta) \neq 0$ implies that $\eta \neq 0$ i.e. then the true vector $\tilde{\beta}$ and its projection $\tilde{\beta}^*$ are indeed proportional. Also in Proposition 3.9 we show how KL projection on a smaller model depends on the true parameters and the covariance structure of the predictors. Moreover, we study projections on both logistic and linear model and and the interplay between them under Rudd's condition. In particular, we prove in Proposition 3.11 that the projection on the logistic model can be recovered, up to proportionality constant, from the projection on the linear model.The last result is potentially useful when ranking variables according to the absolute values of their estimated coefficients as it asserts that such procedure can be based on a linear fit instead of logistic. This is computationally much less consuming. Finally, we consider the normal case and we state an equation for proportionality constant $\eta$ which is a consequence of Stein's lemma.

In Section 4 we construct several examples showing that any interplay between $t$ and $t^*$ is possible i.e. $t^*$ may be proper subset or superset of $t$, moreover $t$ and $t^*$ may have only 0 in common. In the last section we check numerically the properties of projections discussed in the paper and show that proportionality constant $\eta$ can be calculated using Newton-Raphson procedure.

In Section 5 we show among others how KL projection can be numerically computed in the normal case and study in detail a situation in which misspecification results from omission of an active variable in a logistic model. Section 6 concludes the paper.

We consider here the case when potential predictors $X_1, \ldots, X_p$ are random, where $X = (1, X_1, \ldots, X_p)^T$. We stress that no conditions on dimension $p$ of vector of predictors is imposed in the paper. Kullback-Leibler projection $\beta_n^*(\mathbf{X})$ for deterministic experimental matrix $\mathbf{X}$ can be analogously defined which will now depend on $\mathbf{X}$ and sample size

$n$. For relevant discussion see [7], where asymptotic results for ML estimators centred at $\beta_n^*(\mathbf{X})$ are proved and [18] for asymptotic consistency of some selection methods. The studies of random and deterministic design differ; one of the main differences is that $\beta_n^*(\mathbf{X})$ can be approximated by simple adaptation of Iterated Weighted Least Squares algorithm used to calculate ML estimators, whereas $\beta^*$ is given only indirectly as solution to the normal equations discussed below. The area of interplay between $t$ and support of $\beta_n^*(\mathbf{X})$ is equally uncharted but is beyond the objective of this paper.

## 2. Preliminaries

We will use the following notation $X = (1, \tilde{X}^T)^T$ and $\beta = (\beta_0, \tilde{\beta}^T)^T$, where $\tilde{X}$ is a vector of predictors and $\tilde{\beta}$ the corresponding vector of parameters. We discuss first several facts which will be used in the development. It is proved in [16], Lemma 3.1 that $\beta^*$ exists (but it is not necessary unique) provided $E||X|| < \infty$ and

$$P(q(X^T\beta) \in (0,1)) = 1. \tag{9}$$

Moreover, when the second moment of $X$ is finite, changing the order of differentation and averaging we have that

$$\frac{\partial}{\partial b} E(\Delta_X(q(X^T\beta), q_L(X^T b)) = EX(Y - q_L(X^T b)) \tag{10}$$

and

$$H = \frac{\partial^2}{\partial b \partial b^T} E(\Delta_X) = -E(q_L(X^T b)(1 - q_L(X^T b))XX^T). \tag{11}$$

It is easy to observe that Hessian is negative definite if the moment matrix $EXX^T$ exists and is positive definite. In order to see this, note that the positive definiteness of $EXX^T$ implies that $E|X^T\lambda|^2 > 0$ for any $\lambda \in R^p \setminus \{0\}$ and thus $P(A) > 0$, where $A = \{X^T\lambda \neq 0\}$. It follows that $\lambda^T H \lambda = -E(q_L(X^T b)(1 - q_L(X^T b))|X^T\lambda|^2 I_A) < 0$ as $q_L(s) > 0$ for any $s \in R$. Note that the positive definiteness of $EXX^T$ is equivalent to that of $E\tilde{X}\tilde{X}^T$. Thus under this condition and (9), KL projection $\beta^*$ exists and is unique which will be assumed throughout. Note that it follows from (10) that $\beta^*$ satisfies the normal equations

$$E(X(q(X^T\beta) - q_L(X^T\beta^*)) = 0, \tag{12}$$

or equivalently, as the first coordinate of $X$ is $X_0 \equiv 1$, we have

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, q(X^T\beta)) = \mathrm{Cov}(X, q_L(X^T\beta^*)). \tag{13}$$

Due to the structure of the normal equations, their explicit solution, apart from the case discussed below, is rarely known for continuous predictors. Analogously to $\beta = (\beta_0, \tilde{\beta}^T)^T$, we set $\beta^* = (\beta_0^*, \tilde{\beta}^{*T})^T$, moreover let $t = \{i : \tilde{\beta}_i \neq 0\} \cup \{0\}$ and $t^* = \{i : \tilde{\beta}_i^* \neq 0\} \cup \{0\}$ i.e. supports of $\tilde{\beta}$ and $\tilde{\beta}^*$, respectively, with index of intercept included.
In the following we will consider KL projections for models pertaining to subsets of regressors. Recall that $\beta^*$ denotes the parameter of KL projection on all regressors. It is

easy to see that if model $s \subseteq \{1, 2, \ldots, p\}$ is such that $t \cup t^* \subseteq s$ then restricting $\beta^*$ to $s$ yields projection of $\beta$ on the logistic model involving only predictors in $s$. Thus removing unnecessary zeros from vector $\beta^*$ we obtain projections of smaller models. We prove in Theorem 3.1 below that under certain assumptions the opposite is true, namely that $\beta^*$ can be reconstructed from the parameter of projection on model $t$ by appending it with zeros.

We will use Ruud's theorem (cf [21], see also [16]) which says that if regressions of $\tilde{X}$ given $\tilde{X}^T\tilde{\beta}$ are linear, then $\tilde{\beta}^*$ has the same direction as $\tilde{\beta}$. More specifically, if the distribution of $\tilde{X}$ is such that $\tilde{X}^T\tilde{\beta}$ is nondegenerate and the regression $E(\tilde{X}|\tilde{X}^T\tilde{\beta} = z)$ is linear in $z$ i.e.

$$E(\tilde{X}|\tilde{X}^T\tilde{\beta} = z) = uz + u_0 \tag{14}$$

for some $u_0, u \in R^p$. Then Ruud's theorem asserts that there exists $\eta \in R$ such that $\tilde{\beta}^* = \eta\tilde{\beta}$. Note that proportionality holds for vectors of parameters pertaining to $X_1, \ldots, X_p$ excluding intercept.

Condition (14) holds for elliptically contoured distributions, in particular, the normal. Thus, provided $\eta \neq 0$ in such cases true direction of $\tilde{\beta}$ can be approximately recovered by ML estimates. This is of paramount importance in classification. In this context we mention the work [1] of D. Brillinger, who showed that a similar phenomenon holds when fitting a linear regression to a nonlinear one provided the regressors are normal. Moreover, it follows from Rudd's theorem that under condition (14) we have either $t = t^*$ (for $\eta \neq 0$) or $t^* = \{0\}$. In [12] it is observed that if $X$ is standardised, $E\tilde{X} = 0$, $\mathrm{Var}\tilde{X} = I$, and $||\tilde{\beta}|| = 1$ then $u = \beta$. Indeed, it follows from decomposition of $\mathrm{Var}\tilde{X}$ that $I = \mathrm{Cov}(u\tilde{X}^T\tilde{\beta}) + E(\mathrm{Cov}(\tilde{X}|\tilde{X}^T\tilde{\beta}))$ which implies $|\tilde{\beta}^T u| = 1$ and also that $I \geq \mathrm{Cov}(u\tilde{X}^T\tilde{\beta}) = uu^T$. From the inequality it follows that $u^T u \geq (u^T u)^2$ and thus $||u|| \leq 1$. Now the Schwarz inequality together with $|\tilde{\beta}^T u| = 1$ implies $u = \tilde{\beta}$ as $||\tilde{\beta}|| = 1$ and $\tilde{X}^T\tilde{\beta}$ is non-degenerate. The property that $u = \tilde{\beta}$ holds, however, for standardized $\tilde{X}$ only. We also note that (14) implies in view of (2) that

$$E(\tilde{X}|Y) = E(E(\tilde{X}|\tilde{X}^T\tilde{\beta}, \varepsilon)|Y) = E(E(\tilde{X}|\tilde{X}^T\tilde{\beta})|Y) = u_0 + uE(\tilde{X}^T\tilde{\beta}|Y) \tag{15}$$

which is a principal motivation behind sliced inverse regression introduced in [17].

*Remark* 2.1 It follows easily from (12) that $t^* = \{0\}$ is equivalent to $\mathrm{Cov}(Y, \tilde{X}) = 0$ or to $E(\tilde{X}|Y = 1) = E(\tilde{X}|Y = 0)$. Note that the last condition implies $E(\tilde{X}^T\tilde{\beta}|Y = 1) = E(\tilde{X}^T\tilde{\beta}|Y = 0)$ and for the normal $X$ both conditions are equivalent in the view of (15).

Some results on interplay between $t$ and $t^*$ can be obtained when in place of linear regressions' property some specific condition on response $q$ is assumed. In particular, it is proved in Theorem 4 of [19] that if $q$ is monotone and not constant and distribution $P_X$ is not linearly degenerate then $t = \{0\}$ is equivalent to $t^* = \{0\}$. Thus for such response functions we can not overlook dependence of $Y$ on $X$ (i.e. have $t^* = \{0\}$) by misspecifying the model.

## 3. Main results

We first deal with the general case when no specific assumptions on distribution $P_{\tilde{X}}$ of predictors $\tilde{X} = (X_1, \ldots, X_p)^T$ are imposed. We assume throughout that projection $\beta^*$

exists and is unique. The first main result states that when inactive predictors $\tilde{X}_2$ in binary model are such that $\tilde{X}_1$ and $\tilde{X}_2 - A\tilde{X}_1$ are independent, where $\tilde{X}_1$ are remaining predictors and $A$ is a linear transform, then KL projection on $\tilde{X}_1$ and $\tilde{X}_2$ jointly is obtained from KL projection on $\tilde{X}_1$ alone by appending zeroes to the latter. It is easily seen that the result fails when vectors $\tilde{X}_1$ and $\tilde{X}_2$ are dependent; see Example 4.3 below.

### 3.1. *General case*

THEOREM 3.1 *Let $\tilde{X} = (\tilde{X}_1^T, \tilde{X}_2^T)^T$, where $\tilde{X}_1 = (X_1, \ldots, X_k)^T, \tilde{X}_2 = (X_{k+1}, \ldots, X_p)^T$, $\beta = (\beta_0, \tilde{\beta}_1^T, \tilde{\beta}_2^T)^T$ and $\beta^*$ is KL projection of $\beta$. Assume that $\tilde{\beta}_2 = 0$ and let $(\beta_0^*, \tilde{\beta}_1^{*T})^T$ be the KL projection of $(\beta_0, \tilde{\beta}_1^T)^T$. If $\tilde{X}_1$ and $\tilde{X}_2 - A\tilde{X}_1$ are independent for a certain $A \in R^{(p-k)\times k}$, then $\beta^* = (\beta_0^*, \tilde{\beta}_1^{*T}, 0_{p-k}^T)^T$.*

*Proof.* From normal equations (12) for $\tilde{\beta}_1^*$ we have

$$Eq_L(\beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*) = Eq(\beta_0 + \tilde{X}_1^T\tilde{\beta}_1) \tag{16}$$

and

$$Eq_L(\beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*)\tilde{X}_1 = Eq(\beta_0 + \tilde{X}_1^T\tilde{\beta}_1)\tilde{X}_1. \tag{17}$$

Thus it is enough to show that

$$Eq_L(\beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*)\tilde{X}_2 = Eq(\beta_0 + \tilde{X}_1^T\tilde{\beta}_1)\tilde{X}_2. \tag{18}$$

Indeed, observe that

$$\begin{aligned}
& Eq_L(\beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*)\tilde{X}_2 = Eq_L(\beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*)(\tilde{X}_2 - A\tilde{X}_1) + Eq_L(\beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*)A\tilde{X}_1 \\
= \;& Eq_L(\beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*)E(\tilde{X}_2 - A\tilde{X}_1) + AEq_L(\beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*)\tilde{X}_1 \\
= \;& Eq(\beta_0 + \tilde{X}_1^T\tilde{\beta}_1)E(\tilde{X}_2 - A\tilde{X}_1) + AEq(\beta_0 + \tilde{X}_1^T\tilde{\beta}_1)\tilde{X}_1 \\
= \;& Eq(\beta_0 + \tilde{X}_1^T\tilde{\beta}_1)(\tilde{X}_2 - A\tilde{X}_1) + Eq(\beta_0 + \tilde{X}_1^T\tilde{\beta}_1)A\tilde{X}_1 \\
= \;& Eq(\beta_0 + \tilde{X}_1^T\tilde{\beta}_1)\tilde{X}_2,
\end{aligned}$$

where the second line follows from independence of $\tilde{X}_2 - A\tilde{X}_1$ and $\tilde{X}_1$ and linearity of $A$, the third from normal equations (16) and (17) and the fourth again from independence of $\tilde{X}_2 - A\tilde{X}_1$ and $\tilde{X}_1$ and linearity of $A$. Thus it follows that (18) is satisfied and $(\beta_0^*, \tilde{\beta}_1^{*T}, 0_{p-k}^T)^T$ satisfies normal equations for the KL projection on $\beta$. Now the result follows from the uniqueness of $\beta^*$.

■

Let $X_t = \sum_{i=1}^p a_i X_{t-i} + \varepsilon_t$, $a_i \in R, t \in Z$ be a causal autoregressive AR($p$) process where $(\varepsilon_t)$ is a sequence of i.i.d. random variables with finite second moment (see [2], Chapter 3). Let $\tilde{X}_1 = (X_n, \ldots, X_1)^T$ and $\tilde{X}_2 = X_{n+1}$. Then for $n \geq p$ with $A = (a_1, \ldots, a_p, 0_{n-p})$, where $0_{n-p} \in R^{n-p}$ is a vector consisting of zeros, we have that $\tilde{X}_1$ and $\tilde{X}_2 - A\tilde{X}_1 = \varepsilon_{n+1}$ are independent and the assumption of Theorem 3.1 is satisfied.

COROLLARY 3.2 *If we additionally assume that $\beta_i \neq 0$ for all $i \in \{1, \ldots, k\}$ i.e. $t = \{1, \ldots, k\}$ then $t^* \subseteq t$. Moreover, if $q$ is monotone and not constant and distribution $P_{\tilde{X}}$ is not linearly degenerate we have $\{0\} \subset t^* \subseteq t$.*

The second part of the corollary follows from Theorem 4 in [19] discussed in the previous section.

*Example* 3.3 In order to illustrate the conclusion of Theorem 3.1 we consider data set `student` from uci depository (`https://archive.ics.uci.edu/ml/datasets/Student+Performance`). The variables include `G3` and `G2` denoting student's grade respectively at the end of the third and the second semester and binary variable `school`. The range of `G3` and `G2` is [0,20]. We considered binary response $Y = I\{\texttt{G3} \geq 10\}$ and predictors $X_1 = \texttt{G2}$, $X_2 = \texttt{school}$. Two logistic models using function `glm` in `R` were fitted:

$$M1: \quad Y \sim X_1 + X_2,$$

$$M2: \quad Y \sim X_1.$$

In this example predictor $X_2$ is neither correlated with $X_1$ (Pearson correlation $\rho = -0.05$, p-value=0.32) or with $Y$ ($\rho = -0.03$, p-value=0.53) thus the assumptions of Theorem 3.1 are satisfied with $A = 0$. ML estimates were considered as proxies for KL projections on both models. The ML fit for model M1 yields

$$\hat{\beta}_{M1} = (\hat{\beta}_{0,M1}, \hat{\beta}_{1,M1}, \hat{\beta}_{2,M1})^T = (-17.16, 1.87, 0.41)^T$$

with null and residual deviances equal to 500.5 and 148.56, respectively. Analogously, for model M2 we obtain

$$\hat{\beta}_{M2} = (\hat{\beta}_{0,M2}, \hat{\beta}_{1,M2})^T = (-17.06, 1.86)^T$$

with residual deviance equal to 148.94. It turns out using Wald test that $\hat{\beta}_{2,M1}$ is insignificant in model M1. Moreover, values of intercepts and estimators of $\beta_1$ are very close in both models. This suggests that indeed KL projection on M1 is KL projection on M2 appended by 0. In order to additionally confirm this we compute the deviance test comparing the models $M_2$ and $M_1$ which yields

$$\text{dev}_{M2,M1} = 148.94 - 148.56 = 0.38,$$

which is obviously insignificant with p-value=0.96 based on chi-square $\chi^2_1$.

The next result states conditions under which KL projections are continuous with respect to distribution of predictors. The result which is interesting in its own right will also be used to establish the fact that for any response function $q$ there exists continuous random variable supported on the set of non-zero Lebesque measure such that $t$ and $t^*$ have only 0 in common.

THEOREM 3.4 *Assume that $X_n$ and $X$ are random variables such that $E||X_n - X||_2 \to 0$, $X$ is integrable and $q$ is uniformly continuous. For $\beta^*_n$ and $\beta^*$ denoting KL projections for binary models $P(Y_n = 1|X_n) = q(X_n^T\beta)$ and $P(Y = 1|X) = q(X^T\beta)$, respectively, we have $\beta^*_n \to \beta^*$.*

*Proof.* Let $l_n(b) = l(b, X_n, Y_n), l(b) = l(b, X, Y)$. We first note that the uniform convergence holds for $|El_n(b) - El(b)|$ on bounded sets i.e. for any finite $K$

$$\sup_{||b||\leq K} |El_n(b) - El(b)| \to 0. \tag{19}$$

Indeed, using the Schwarz inequality, the mean value theorem and boundedness of $q$ we get the following sequence of inequalities:

$$\begin{aligned} & |El_n(b) - El(b)| \leq |E(q(X_n^T\beta)X_n^Tb - q(X^T\beta)X^Tb)| + ||b||_2E||X_n - X||_2 \\ \leq\ & |E(q(X_n^T\beta) - q(X^T\beta))X^Tb| + E|q(X_n^T\beta)||X_n^Tb - X^Tb| + ||b||_2E||X_n - X||_2 \\ \leq\ & E|(q(X_n^T\beta) - q(X^T\beta))X^Tb| + 2||b||_2E||X_n - X||_2. \end{aligned}$$

Choosing for arbitrary $\varepsilon > 0$ $\delta > 0$ such that for large $n$ if $||X_n - X||_2 < \delta$, then $|q(X_n^T\beta) - q(X^T\beta)| < \varepsilon$, we have

$$\begin{aligned} & E|(q(X_n^T\beta) - q(X^T\beta))X^Tb| \leq \varepsilon EI(||X_n - X||_2 < \delta)|X^Tb| + EI(||X_n - X||_2 \geq \delta)|X^Tb| \\ \leq\ & \varepsilon||b||_2E||X||_2 + ||b||_2EI(||X_n - X||_2 \geq \delta)||X||_2 \leq C\varepsilon||b||_2, \end{aligned}$$

for large $n$, where $C$ is some constant. From this convergence in (19) readily follows as $\varepsilon > 0$ was arbitrary. We now prove that $\beta_n^* \to \beta^*$. If does not hold then for a certain $k_n \in N$ tending to infinity we have $|\beta_{k_n}^* - \beta^*| \geq \delta$ for some $\delta > 0$. From Lemma 2 in [14] and uniqueness of $\beta^*$ it follows that

$$\sup_{||b-\beta^*||\leq\delta} |El_{k_n}(b) - El(b)| \geq \frac{1}{2}(El(\beta^*) - \sup_{||b-\beta^*||=\delta} El(b)) > 0.$$

which contradicts (19). ■

We now prove a negative result showing that for any response function $q$, supports $t$ and $t^*$ may have only 0 in common for a certain distribution $X$. We let below $t = t(X)$ and $t^* = t^*(X)$ to stress the dependence of both sets on distribution of $X$.
To show that such random variable exists, we will first state the following two lemmas, the proofs of which are relegated to the Appendix.

LEMMA 3.5 *Let* $Z_m = (Z_{m1}, \ldots, Z_{mp})^T \sim \sum_{l=1}^{p} p_l\mathcal{N}_p\left(x_l, \frac{\sigma^2}{m}I_p\right), p_l > 0, \sum_{l=1}^{n} p_l = 1, x_l \in R^p, \sigma > 0$. *Let* $P(Z = x_l) = p_l, l = 1, \ldots, p$. *If* $t^*(Z) = \{0, 1, \ldots, p\}$, *then* $t^*(Z_m) = t^*(Z)$ *for sufficiently large* $m$ *and uniformly continuous function* $q$.

We say that $Z$ is linearly non-degenerate variable if for all $b \in R^p, c \in R$ we have $P(b^TZ = c) < 1$.

LEMMA 3.6 *For any* $p \in N$, $p \geq 2$, $k \in \{1, \ldots, p-1\}$ *and* $q$ *being continuous response function such that for all* $a, b \in R$ *there exists* $x \in R$ $q(x) \neq q_L(ax + b)$ *there exists* $Z$ *such that* $t^*(Z) = \{0, 1, \ldots, p\}$, $t(Z) = \{0, 1, \ldots, k\}$ *and* $Z$ *is linearly nondegenerate.*

THEOREM 3.7 *For any uniformly continuous response function* $q$ *there exists* $R^p$*-valued random variable* $X$ *supported on the set of non-zero Lebesque measure and such that* $t(X) \cap t^*(X) = \{0\}$.

*Proof.* In order to prove the theorem we apply Lemma 3.5 to a discrete variable $Z$ constructed as in Lemma 3.6 and $\beta_i = I\{i \leq k\},\ i = 1, \ldots, p$. Let $Z_m$ from Lemma 3.5 for sufficiently large $m$ be such that $t^*(Z_m) = \{0, 1, \ldots, p\}$. From the construction $t(Z_m) = \{0, 1, \ldots, k\}$. Let $X_i = Z_{mi}$ for $i \leq k$, where $Z_{mi}$ is defined in Lemma 3.5, $X_{k+1} = \sum_{i=1}^{k+1} \beta_i^*(Z_m) Z_{mi}$, $X_{k+1+i} = \beta_{k+1+i}^*(Z_m) Z_{m,k+1+i}$ for every $p-1-k \geq i > 0$. Then we show that $t(X) = \{0, 1, \ldots, k\}$, $t^*(X) = \{0, k+1, \ldots, p\}$, i.e. $t(X) \cap t^*(X) = \{0\}$. Indeed, normal equations for the vector $Z_m$ have the form

$$Eq_L(\beta^{*T}(Z_m)Z_m)Z_m = Eq(\beta^T(Z_m)Z_m)Z_m = Eq\left(\sum_{i=1}^{k} \beta_i(Z_m)Z_{mi}\right) Z_m.$$

By rewriting them for vector $X$, we obtain:

$$Eq_L\left(\sum_{i=k+1}^{p} X_i\right) X = Eq\left(\sum_{i=1}^{k} \beta_i(Z_m)X_i\right) X.$$

We can easily see that $t(X) = \{0, 1, \ldots, k\}$. In turn from the uniqueness of projection we obtain $t^*(X) = \{0, k+1, \ldots, p\}$.

■

### 3.2. *Predictors satisfying Ruud's condition*

We will say that Ruud's condition holds for random vector $\tilde{X} = (X_1, \ldots, X_p)^T$ and $\tilde{\gamma} \in R^p$ if the analogue of equation (14) holds, namely

$$E(\tilde{X}|\tilde{X}^T\tilde{\gamma}) = u\tilde{X}^T\tilde{\gamma} + u_0 \tag{20}$$

for some $u_0, u \in R^p$. The above condition means that for any $i = 1, \ldots, p$ regression $E(X_i|\tilde{X}^T\tilde{\gamma} = z)$ is a linear function of $z$. Thus $E(X_i|\tilde{X}^T\tilde{\gamma} = z)$ and $E(X_j|\tilde{X}^T\tilde{\gamma} = z)$ for $i \neq j$ are two lines having in general different slopes and intercepts determined by coordinates of $u$ and $u_0$, respectively. Condition (20) is satisfied in particular for all $\tilde{\gamma} \in R^p$ by multivariate normal predictors or, more generally, the ones having an eliptically contoured distribution.

We now deal with such predictors and show that some results on the form of KL projections can be sharpened in this case. We start with a variant of Ruud's theorem. We show that under a little bit stronger assumption than (14), vectors $\tilde{\beta}$ and $\tilde{\beta}^*$ satisfy equation (21) below from which their proportionality follows and which implies condition $\mathrm{Cov}(Y, X^T\beta) \neq 0$ for proportionality constant $\eta$ to be non-zero. Let

$$a_\gamma = \frac{\mathrm{Cov}(X^T\gamma, Y)}{\mathrm{Var}(X^T\gamma)}$$

for $\gamma \neq 0$ and $a_\gamma = 0$ for $\gamma = 0$. Note that

$$a_\beta = \mathrm{Cov}(X^T\beta, q(X^T\beta))/\mathrm{Var}(X^T\beta)$$

and in view of (13)

$$a_{\beta^*} = \mathrm{Cov}(X^T\beta^*, q_L(X^T\beta^*))/\mathrm{Var}(X^T\beta^*).$$

Moreover, let $\Sigma = \mathrm{Var}(\tilde{X})$. Proposition below asserts proportionality of vectors $\tilde{\beta}$ and $\tilde{\beta}^*$. Note that intercepts $\tilde{\beta}_0$ and $\tilde{\beta}_0^*$ are excluded from equation (21).

PROPOSITION 3.8 *Assume that $\tilde{X}$ and $\tilde{\gamma} = \tilde{\beta}, \tilde{\beta}^*$ satisfy (20), $\Sigma$ is invertible, and moreover $E||X||^2$ is finite. Then*

$$\tilde{\beta} a_\beta = \tilde{\beta}^* a_{\beta^*}. \tag{21}$$

*Further, if $\mathrm{Cov}(Y, X^T\beta) \neq 0$ then then proportionality constant $\eta$ such that $\tilde{\beta}^* = \eta\tilde{\beta}$ satisfies*

$$\eta = a_\beta / a_{\beta^*} \neq 0. \tag{22}$$

*Proof.* We use generalization of Lemma 1 in [1] which states that for bivariate normal $(U, V)$ and measurable $g$ we have that $\mathrm{Cov}(g(U), V) = \mathrm{Cov}(U, V)\mathrm{Cov}(g(U), U)/\mathrm{Var}(U)$ provided that the covariances exist. Namely, it holds that for $i = 1, \ldots, p$

$$\mathrm{Cov}(q(X^T\beta), X_i) = \mathrm{Cov}(X^T\beta, X_i)\mathrm{Cov}(q(X^T\beta), X^T\beta)/\mathrm{Var}(X^T\beta) \tag{23}$$

and

$$\mathrm{Cov}(q_L(X^T\beta^*), X_i) = \mathrm{Cov}(X^T\beta^*, X_i)\mathrm{Cov}(q_L(X^T\beta^*), X^T\beta^*)/\mathrm{Var}(X^T\beta^*). \tag{24}$$

This follows as in [1] after noting that only linearity of conditional expectations $E(\tilde{X}|\tilde{X}^T\tilde{\beta})$ and $E(\tilde{X}|\tilde{X}^T\tilde{\beta}^*)$ is needed respectively for (23) and (24) to hold. Thus using (23) we have $\mathrm{Cov}(q(X^T\beta), X_i) = a_\beta \Sigma_{(i)}\tilde{\beta}$ for $i = 1, \ldots, p$, where $\Sigma_{(i)}$ is $i^{\text{th}}$ row of $\Sigma$. Analogously, using the same reasoning for all coordinates of the right hand side of (24) we see that it follows from (13) that

$$\Sigma\tilde{\beta} a_\beta = \Sigma\tilde{\beta}^* a_{\beta^*}. \tag{25}$$

As $\Sigma$ is invertible, the first part of the proposition follows. The second one is implied by the equality $\mathrm{Cov}(Y, X^T\beta) = \mathrm{Cov}(q(X^T\beta), X^T\beta) \neq 0$. Thus $a_{\beta^*}$ is nonzero and (22) holds. ■

Note that for strictly increasing, positive valued real function $f$ and $\beta \neq 0$ we have, provided $E|f(X^T\beta)X^T\beta|$ is finite, that

$$\mathrm{Cov}(f(X^T\beta), X^T\beta) > 0.$$

Whence if $q$ is strictly increasing and positive and conditions of Proposition 3.8 are satisfied, then $\eta > 0$ and $\tilde{\beta}^*$ has the same direction and *orientation* as $\tilde{\beta}$.
We consider now a more general setting for which in addition to misspecification of response function $q$ we assume that only subvector $\tilde{X}_1$ of all predictors in (1) is fitted. We show that under analogous assumptions to those of Proposition 3.8 KL projections

on a smaller model corresponding to a subset of predictors can be determined up to a proportionality constant.

PROPOSITION 3.9 *Let* $\tilde{X} = (\tilde{X}_1^T, \tilde{X}_2^T)^T$, $\beta = (\beta_0, \tilde{\beta}_1^T, \tilde{\beta}_2^T)^T$ *and* $\tilde{\beta}_1, \tilde{X}_1 \in R^m, \tilde{\beta}_2, \tilde{X}_2 \in R^{p-m}$, $\mathrm{Cov}(\tilde{X}_i, \tilde{X}_j) = \Sigma_{ij}$ *for* $i, j = 1, 2$. *Suppose that logistic model* $Y \sim \beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*$ *with omitted* $\tilde{X}_2$ *variables is KL projection of (1). Under assumptions of Proposition 3.8 for* $\tilde{X}$ *and* $\tilde{\gamma} = \tilde{\beta}, \tilde{\beta}_1^*$ *and provided that* $\mathrm{Cov}(Y, \tilde{X}_1^T\tilde{\beta}_1) \neq 0$ *we have*

$$\tilde{\beta}_1^* = \eta(\tilde{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\tilde{\beta}_2), \tag{26}$$

*where* $\eta = a_\beta / a_{\beta_1^*} \neq 0$ *and*

$$a_{\beta_1^*} = \frac{\mathrm{Cov}(Y, \tilde{X}_1^T\tilde{\beta}_1^*)}{\mathrm{Var}(\tilde{X}_1^T\tilde{\beta}_1^*)} = \frac{\mathrm{Cov}(q_L(\tilde{\beta}_0^* + \tilde{X}_1^T\tilde{\beta}_1^*), \tilde{X}_1^T\tilde{\beta}_1^*)}{\mathrm{Var}(\tilde{X}_1^T\tilde{\beta}_1^*)}. \tag{27}$$

*Proof.* Analogously as in Proposition 3.8, we obtain the equations:

$$\mathrm{Cov}(q(\beta_0 + \tilde{X}_1^T\tilde{\beta}_1 + \tilde{X}_2^T\tilde{\beta}_2), \tilde{X}_1) = a_\beta \mathrm{Cov}(\tilde{X}_1, \tilde{X}_1^T\tilde{\beta}_1 + \tilde{X}_2^T\tilde{\beta}_2),$$

$$\mathrm{Cov}(q_L(\beta_0^* + \tilde{X}_1^T\tilde{\beta}_1^*), \tilde{X}_1) = a_{\beta_1^*} \mathrm{Cov}(\tilde{X}_1, \tilde{X}_1^T\tilde{\beta}_1^*).$$

Hence from normal equations we get

$$a_{\beta_1^*} \mathrm{Cov}(\tilde{X}_1, \tilde{X}_1^T\tilde{\beta}_1^*) = a_\beta \mathrm{Cov}(\tilde{X}_1, \tilde{X}_1^T\tilde{\beta}_1 + \tilde{X}_2^T\tilde{\beta}_2),$$

which can be simplified to

$$a_{\beta_1^*}\Sigma_{11}\tilde{\beta}_1^* = a_\beta(\Sigma_{11}\tilde{\beta}_1 + \Sigma_{12}\tilde{\beta}_2). \tag{28}$$

As $\mathrm{Cov}(\tilde{X})$ is invertible thus $\Sigma_{11}$ is invertible and similarly as in Proposition 3.8 we conclude that $a_{\beta_1^*}$ is nonzero. Multiplying both sides of (28) by $(a_{\beta_1^*}\Sigma_{11})^{-1}$, we obtain the conclusion. ■

*Remark* 3.10 If in Proposition 3.9 we assume additionally that $\Sigma_{12} = 0$ then $\beta_1^* = \eta\beta_1$. For independent $\tilde{X}_1$ and $\tilde{X}_2$ we thus obtain a complemetary conclusion to that of Theorem 3.1.

We consider now an analogous problem of how the coefficients of a linear and logistic regression fitted to binary model (1) compare. Suppose that logistic and linear model with $\tilde{X}_1$ predictors are fitted to binary model (1), where $\tilde{X} = (\tilde{X}_1^T, \tilde{X}_2^T)^T$. Thus we omit variables $\tilde{X}_2$ from the fit. In order to stress that two different models are fitted we denote $\beta_{1,log}^* = (\beta_{0,log}^*, \tilde{\beta}_{1,log}^{*T})^T$ parameters of KL projection on the logistic model and by $\beta_{1,lin}^* = (\beta_{0,lin}^*, \tilde{\beta}_{1,lin}^{*T})^T$ parameters for the linear model. We rename accordingly $a_{\beta_1^*}$ defined in (27) as $a_{\beta_{1,log}^*}$. Let $\mathrm{Var}(\tilde{X}_1) = \Sigma_{\tilde{X}_1}$. We show that projections on logistic and linear model are proportional. Namely, we have

PROPOSITION 3.11 *Assume that Ruud's condition holds for $\tilde{X}_1$ and $\beta^*_{1,log}$ and $\Sigma_{\tilde{X}_1}$ is invertible. Then*

$$\tilde{\beta}^*_{1,lin} = a_{\beta^*_{1,log}} \tilde{\beta}^*_{1,log}. \tag{29}$$

*Proof.* For any $i$ being an index of coordinate of $\tilde{X}_1$ we have

$$\mathrm{Cov}(q_L(\beta^*_{0,log} + \tilde{\beta}^{*T}_{1,log}\tilde{X}), X_i) = \mathrm{Cov}(Y, X_i) = \mathrm{Cov}(\beta^*_{0,lin} + \tilde{\beta}^{*T}_{1,lin}\tilde{X}, X_i) = \tilde{\beta}^{*T}_{1,lin}\Sigma^{(i)}_{\tilde{X}_1},$$

where $\Sigma^{(i)}_{\tilde{X}_1}$ is $i^{th}$ column of covariance matrix of $\tilde{X}_1$. Moreover, reasoning as in proof of Proposition 3.8 we have for such $i$

$$\mathrm{Cov}(q_L(\beta^*_{0,log} + \tilde{\beta}^{*T}_{1,log}\tilde{X}), X_i) = a_{\beta^*_{1,log}} \cdot \tilde{\beta}^{*T}_{1,log}\Sigma^{(i)}_{\tilde{X}_1}.$$

Thus from two last equalities we obtain matrix equation

$$a_{\beta^*_{1,log}}\Sigma_{\tilde{X}_1}\tilde{\beta}^*_{1,log} = \Sigma_{\tilde{X}_1}\tilde{\beta}^*_{1,lin},$$

which is equivalent to (29). ■

*Remark* 3.12 In particular the result hold for $\tilde{X}_1 = \tilde{X}$ when all regressors are fitted and for $\tilde{X}_1 = X_j$ when the univariate regressor $X_j$ is fitted. In the latter case Ruud's condition for $X_j$ and $\beta^*_{1,log}$ is always satisfied. Note also that when $q \equiv q_L$ and $\tilde{X}_1 = \tilde{X}$ i.e. the model is correctly specified it follows that $\tilde{\beta}^*_{lin}$ is proportional to $\tilde{\beta}$. Thus in this case an important problem of ranking unknown coefficients of logistic model can be based on a fit of a linear model which is much easier computationally than a logistic fit.

*Remark* 3.13 Note that it follows from the last Proposition that in the case $\tilde{X}_1 = \tilde{X}$, $\tilde{\beta}^*_{log}$ can be computed, up to a constant of proportionality, from vector $\bar{\beta}^*_{lin}$ consisting of coefficients of univariate linear filters and covariance matrix $\Sigma$. Namely we have

$$a_{\beta^*} \cdot \Sigma\tilde{\beta}^*_{log} = D\bar{\beta}^*_{lin}, \tag{30}$$

where $D = \mathrm{diag}(\Sigma)$, or equivalently

$$\tilde{\beta}^*_{log} = {a^*_\beta}^{-1}\Sigma^{-1}D\bar{\beta}^*_{lin}. \tag{31}$$

**Normal case** Consider now the case when $X$ is multivariate normal. In this case important additional conclusion can be inferred from Stein's lemma in [22]. Namely, it follows from it that for differentiable $q$ such that $E|q'(X^T\beta)|$ exists we have that

$$a_\beta = Eq'(X^T\beta).$$

Moreover

$$a_{\beta^*} = Eq'_L(X^T\beta^*).$$

Thus it holds for $\eta$ from Proposition 3.8 that

$$\eta = \frac{Eq'(X^T\beta)}{Eq'_L(X^T\beta^*)} = \frac{Eq'(X^T\beta)}{Eq'_L(\eta X^T\beta)}, \tag{32}$$

which yields an equation satisfied by $\eta$. Moreover, the conclusion of Proposition 3.11 can be stated as

$$\tilde{\beta}^*_{1,lin} = Eq'_L(\beta^*_{0,log} + \tilde{\beta}^*_{1,log}\tilde{X}_1)\tilde{\beta}^*_{1,log}.$$

Observe that it follows in particular from the last equality that $4||\tilde{\beta}^*_{1,lin}|| \leq ||\tilde{\beta}^*_{1,log}||$ as $q'_L(s) = q_L(s)(1-q_L(s))$ and $0 \leq t(1-t) \leq 1/4$ for $t \in [0,1]$ .

## 4. Examples

In this section we provide several examples showing that it may happen that $t^*$ is a proper subset or superset of $t$. Also we provide constructive examples of the situation when $t \cap t^* = \{0\}$. Note that although in the first three examples predictors are discrete, we can have the same properties for continuous predictors $\tilde{X}$ by replacing discrete $\tilde{X}$s constructed below with mixtures of normally distributed variables centred at respective atoms (cf Lemma 3.5).

*Example* 4.1 $t^* \subset t$
Let vector $(X_1, X_2)$ have the distribution:
$P(X_1 = 1, X_2 = 0) = P(X_1 = X_2 = 0) = P(X_1 = X_2 = 2) = \frac{1}{3}$,
and
$q(x) = \begin{cases} q_L(x) \text{ if } x \leq 1 \\ q_L(\sqrt{x}) \text{ otherwise.} \end{cases}$
Let $\beta_0 = 0, \beta_1 = \beta_2 = 1$.
The normal equations are:
$\begin{cases} Eq_L(\beta^*_0 + \beta^*_1 X_1 + \beta^*_2 X_2) = Eq(X_1 + X_2) \\ Eq_L(\beta^*_0 + \beta^*_1 X_1 + \beta^*_2 X_2)X_1 = Eq(X_1 + X_2)X_1 \\ Eq_L(\beta^*_0 + \beta^*_1 X_1 + \beta^*_2 X_2)X_2 = Eq(X_1 + X_2)X_2 \end{cases}$
which simplify here to the form:
$\begin{cases} \frac{q_L(\beta^*_0 + \beta^*_1)}{3} + \frac{q_L(\beta^*_0)}{3} + \frac{q_L(\beta^*_0 + 2\beta^*_1 + 2\beta^*_2)}{3} = \frac{q(1)}{3} + \frac{q(0)}{3} + \frac{q(4)}{3} \\ \frac{q_L(\beta^*_0 + \beta^*_1)}{3} + \frac{2q_L(\beta^*_0 + 2\beta^*_1 + 2\beta^*_2)}{3} = \frac{q(1)}{3} + \frac{2q(4)}{3} \\ \frac{2q_L(\beta^*_0 + 2\beta^*_1 + 2\beta^*_2)}{3} = \frac{2q(4)}{3} \end{cases}$
Invertibility of $q_L$ and the form of function $q$ yields:
$\begin{cases} \beta^*_0 = 0 \\ \beta^*_0 + \beta^*_1 = 1 \\ \beta^*_0 + 2\beta^*_1 + 2\beta^*_2 = 2 \end{cases}$
Hence $\beta^*_0 = \beta^*_2 = 0, \beta^*_1 = 1$. This means that $t^* = \{0,1\} \subsetneq \{0,1,2\} = t$.

*Example* 4.2 $t^* \supset t$
Let vector $(X_1, X_2)$ have the distribution such that:

$P(X_1 = 0, X_2 = 1) = P(X_1 = 1, X_2 = 1) = P(X_1 = 2, X_2 = 3) = \frac{1}{3}$
and $q(x) = q_L(x|x|)$.
Moreover, take $\beta_0 = 0, \beta_1 = 1, \beta_2 = 0$.
The normal equations are:
$$\begin{cases} Eq_L(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2) = Eq(X_1) \\ Eq_L(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2)X_1 = Eq(X_1)X_1 \\ Eq_L(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2)X_2 = Eq(X_1)X_2 \end{cases}$$
and they reduce to:
$$\begin{cases} q_L(\beta_0^* + \beta_2^*) + q_L(\beta_0^* + \beta_1^* + \beta_2^*) + q_L(\beta_0^* + 2\beta_1^* + 3\beta_2^*) = q(0) + q(1) + q(2) \\ q_L(\beta_0^* + \beta_1^* + \beta_2^*) + 2q_L(\beta_0^* + 2\beta_1^* + 3\beta_2^*) = q(1) + 2q(2) \\ q_L(\beta_0^* + \beta_2^*) + q_L(\beta_0^* + \beta_1^* + \beta_2^*) + 3q_L(\beta_0^* + 2\beta_1^* + 3\beta_2^*) = q(0) + q(1) + 3q(2) \end{cases}.$$
Hence after simple transformations we get:
$$\begin{cases} q_L(\beta_0^* + \beta_2^*) = q(0) \\ q_L(\beta_0^* + \beta_1^* + \beta_2^*) = q(1) \\ q_L(\beta_0^* + 2\beta_1^* + 3\beta_2^*) = q(2) \end{cases}$$
Invertibility of $q_L$ the form of function $q$ yields:

$$\begin{cases} \beta_0^* + \beta_2^* = 0 \\ \beta_0^* + \beta_1^* + \beta_2^* = 1 \\ \beta_0^* + 2\beta_1^* + 3\beta_2^* = 4 \end{cases}.$$
Hence $\beta_0^* = -1, \beta_1^* = \beta_2^* = 1$ and thus $t^* = \{0, 1, 2\} \supset \{0, 1\} = t$.

*Example* 4.3 $t^* \cap t = \{0\}$
Let vector $(X_1, X_2)$ have the distribution:
$P(X_1 = 0, X_2 = 0) = P(X_1 = 1, X_2 = 1) = P(X_1 = 7, X_2 = 5) = \frac{1}{3}$
and
$$q(x) = \begin{cases} q_L(x) \text{ if } x \leq 1 \\ q_L\left(\frac{2}{3}x + \frac{1}{3}\right) \text{ otherwise.} \end{cases}.$$
Let $\beta_0 = 0, \beta_1 = 1, \beta_2 = 0$.
The normal equations are:
$$\begin{cases} Eq_L(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2) = Eq(X_1) \\ Eq_L(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2)X_1 = Eq(X_1)X_1 \\ Eq_L(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2)X_2 = Eq(X_1)X_2 \end{cases}$$
which simplify to the form :
$$\begin{cases} q_L(\beta_0^*) + q_L(\beta_0^* + \beta_1^* + \beta_2^*) + q_L(\beta_0^* + 7\beta_1^* + 5\beta_2^*) = q(0) + q(1) + q(7) \\ q_L(\beta_0^* + \beta_1^* + \beta_2^*) + 7q_L(\beta_0^* + 7\beta_1^* + 5\beta_2^*) = q(1) + 7q(7) \\ q_L(\beta_0^* + \beta_1^* + \beta_2^*) + 5q_L(\beta_0^* + 7\beta_1^* + 5\beta_2^*) = q(1) + 5q(7) \end{cases}.$$
Hence after simple transformations we obtain:
$$\begin{cases} q_L(\beta_0^*) = q(0) \\ q_L(\beta_0^* + 7\beta_1^* + 5\beta_2^*) = q(7) \\ q_L(\beta_0^* + \beta_1^* + \beta_2^*) = q(1) \end{cases}$$
Similarily as before we get:

$$\begin{cases} \beta_0^* = 0 \\ \beta_0^* + 7\beta_1^* + 5\beta_2^* = 5 \\ \beta_0^* + \beta_1^* + \beta_2^* = 1 \end{cases}.$$

Hence $\beta_2^* = 1, \beta_1^* = \beta_0^* = 0$. It means that sets $t^* = \{0, 2\}$ and $t = \{0, 1\}$ satisfy $t^* \cap t = \{0\}$.

Let us note that this example also shows in connection with Theorem 3.1 that if predictors $\tilde{X}_1$ and $\tilde{X}_2$ there are dependent then KL projection on $\tilde{X} = (\tilde{X}_1^T, \tilde{X}_1^T)^T$ is not always obtained by appending zeros to KL projection on $\tilde{X}_1$ even though $\tilde{\beta}_2 = 0$.

*Example* 4.4 $t^* \cap t = \{0\}$ for continuous $X$
Let $q(x) = q(-x), X_1, \varepsilon \sim \mathcal{U}[-1, 1]$ be independent, $X_2 = k(X_1 + l\varepsilon)^2$ for some arbitrary non-zero constants $k, l$. If $\beta_1 = 1, \beta_2 = \beta_0 = 0$, then from symmetry of distribution and $q$ it follows that $\beta_1^* = 0$. Moreover, if $\mathrm{Cov}(Y, X_2) = \mathrm{Cov}(q(X_1), X_2) \neq 0$, then $\beta_2^* \neq 0$.

Firstly, let us observe that:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ k(X_1 + l\varepsilon)^2 \end{bmatrix} \stackrel{d}{=} \begin{bmatrix} -X_1 \\ k(-X_1 + l\varepsilon)^2 \end{bmatrix} \stackrel{d}{=} \begin{bmatrix} -X_1 \\ k(-X_1 - l\varepsilon)^2 \end{bmatrix} = \begin{bmatrix} -X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

Let $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)^T$, be KL projection of $\beta_1 = 1, \beta_2 = \beta_0 = 0$ in fitted logistic model and $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2)^T$ be KL projection for $\beta_1 = -1, \beta_2 = \beta_0 = 0$. Since distributions of $(X_1, X_2)$ and $(-X_1, X_2)$ coincide it easily follows from normal equations that $\tilde{\beta}$ that $\tilde{\beta}_0 = \beta_0^*$, $\tilde{\beta}_1 = -\beta_1^*$, $\tilde{\beta}_2 = \beta_2^*$. On the other hand, symmetry of $q$ implies that $q(X_1) = q(-X_1)$, hence from uniqueness of projection we have $\beta^* = \tilde{\beta}$. This means that $\beta_1^* = 0$. Suppose now that $\beta_2^* = 0$. Normal equations take the form:

$$\begin{cases} Eq(X_1) = q_L(\beta_0^*) \\ Eq(X_1)X_1 = q_L(\beta_0^*)EX_1 \\ Eq(X_1)X_2 = q_L(\beta_0^*)EX_2 \end{cases}.$$

Note that the second equation is always satisfied, because from symmetry of distribution $X_1$ and function $q$ we get $EX_1 = 0$ and $Eq(X_1)X_1 = Eq(-X_1)(-X_1) = Eq(X_1)(-X_1) = 0$. By replacing $q_L(\beta_0^*)$ in the third equation above with $Eq(X_1)$, we obtain:

$$Eq(X_1)X_2 = q_L(\beta_0^*)EX_2 = Eq(X_1)EX_2.$$

This means that $\mathrm{Cov}(q(X_1), X_2) = 0$, contradicting the assumptions, thus $\beta_2^* \neq 0$.
Figure 1 shows direction of ML estimate for such model when $k = 2$, $l = 0.25$ and $q(s) = \frac{3}{4} - \frac{s^2}{2}$, $s \in [-1, 1]$.
Intuitively, occurrence of one class with large positive and negative $x_1$ forces $\beta_1^*$ to be close to 0 whereas classes of projected points on the second coordinate are linearly separated which suggests that $\beta_2^*$ is significantly different from 0.

## 5. Numerical experiments

In the numerical experiments we considered calculation of Ruud's proportionality constant, assess the effect of omitting valid predictor in a logistic regression model and check numerically equality (31).

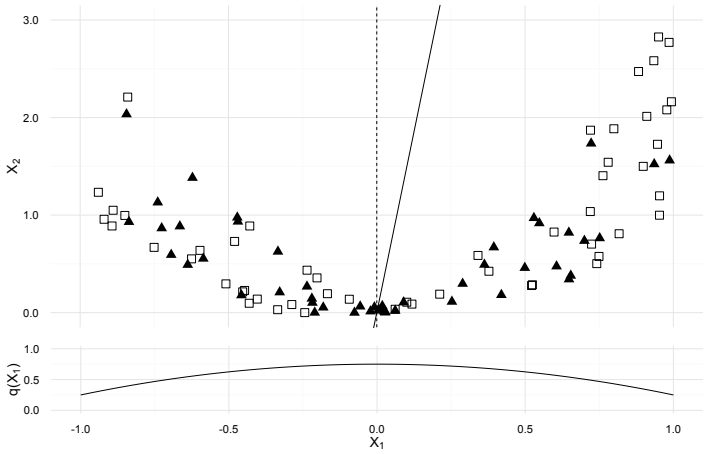



Figure 1. Scatterplot pertaining to the distribution in Example 4. Squares and triangles correspond to $Y = 0$ and $Y = 1$, respectively. Solid line shows the direction of $\hat{\beta}$. The form of $q$ is depicted in the lower plot.

### 5.1. *Calculation of Ruud's proportionality constant by Newton-Raphson procedure*

We mentioned before that finding explicit form of KL projections is rarely possible in continuous case. Nevertheless, our first objective here will be to show that for a given distribution of r.v. $X$, function $q$ and vector $\beta$ computation of KL projection $\beta^*$ is numerically feasible.

We assume that $X$ satisfies Ruud's condition and is linearly non-degenerate. Then we know that $\beta^* = (\beta_0^*, \eta\tilde{\beta}^T)^T$ for some $\eta \in R$. This means that in order to calculate $\beta^*$ it is enough to compute parameters $\beta_0^*$ and $\eta$. In view of normal equation (12) the parameters satisfy $F(\beta_0^*, \eta) = 0$, where

$$F(x,y) = \begin{bmatrix} Eq_L(x + y\tilde{X}^T\tilde{\beta}) - Eq(\beta_0 + \tilde{X}^T\tilde{\beta}) \\ Eq_L(x + y\tilde{X}^T\tilde{\beta})\tilde{X}^T\tilde{\beta} - Eq(\beta_0 + \tilde{X}^T\tilde{\beta})\tilde{X}^T\tilde{\beta} \end{bmatrix}.$$

It is easy to compute the matrix of the first derivatives of $F(x, y)$:

$$J_F(x,y) = \begin{bmatrix} Eq_L'(x + y\tilde{X}^T\tilde{\beta}) & Eq_L'(x + y\tilde{X}^T\tilde{\beta})\tilde{X}^T\tilde{\beta} \\ Eq_L'(x + y\tilde{X}^T\tilde{\beta})\tilde{X}^T\tilde{\beta} & Eq_L'(x + y\tilde{X}^T\tilde{\beta})(\tilde{X}^T\tilde{\beta})^2 \end{bmatrix}.$$

For every $x, y \in R$ we have $\gamma^T J_F(x,y)\gamma > 0$ for $\gamma \in R^2\backslash\{0\}$, whence $J_F(x,y)$ is invertible. The iteration of Newton-Raphson method is thus given by:

$$\begin{bmatrix} \beta_{0,n+1}^* \\ \eta_{n+1} \end{bmatrix} = \begin{bmatrix} \beta_{0,n}^* \\ \eta_n \end{bmatrix} - J_F^{-1}(\beta_{0,n}^*, \eta_n)F(\beta_{0,n}^*, \eta_n).$$

In order to choose a starting point of Newton-Raphson procedure in case of $\tilde{X} \sim \mathcal{N}(0, \Sigma)$, the following approximations can be used

$$Eq(\beta_0 + \tilde{X}^T\tilde{\beta}) = Eq_L(\beta_0^* + \eta\tilde{X}^T\tilde{\beta}) \approx q_L(\beta_0^*)$$

and using Stein's lemma:

$$\eta = \frac{Eq'(\beta_0 + \tilde{X}^T\tilde{\beta})}{Eq'_L(\beta_0^* + \eta\tilde{X}^T\tilde{\beta})} \approx \frac{Eq'(\beta_0 + \tilde{X}^T\tilde{\beta})}{q'_L(\beta_0^*)}.$$

Hence we can take:

$$\beta_0^{*(0)} = q_L^{-1}(Eq(\beta_0 + \tilde{X}^T\tilde{\beta})),$$

$$\eta^{(0)} = \frac{Eq'(\beta_0 + \tilde{X}^T\tilde{\beta})}{q'_L(\beta_0^{*(0)})} = \frac{Eq'(\beta_0 + \tilde{X}^T\tilde{\beta})}{Eq(\beta_0 + \tilde{X}^T\tilde{\beta})(1 - Eq(\beta_0 + \tilde{X}^T\tilde{\beta}))},$$

as $q'_L(\beta_0^{*(0)}) = q_L(\beta_0^{*(0)})(1 - q_L(\beta_0^{*(0)})) = Eq(\beta_0 + \tilde{X}^T\tilde{\beta})(1 - Eq(\beta_0 + \tilde{X}^T\tilde{\beta}))$.
In order to check how the approximation works we considered a setting of numerical experiments in [19]. The following models have been analysed there ($\beta_t$ equals $\beta$ with zeros omitted)

(M1) $t = \{10\}$, $\boldsymbol{\beta}_t = 0.2$,
(M2) $t = \{2, 4, 5\}$, $\boldsymbol{\beta}_t = (1, 1, 1)'$,
(M3) $t = \{1, 2\}$, $\boldsymbol{\beta}_t = (0.5, 0.7)'$,
(M4) $t = \{1, 2\}$, $\boldsymbol{\beta}_t = (0.3, 0.5)'$,
(M5) $t = \{1, \ldots, 8\}$, $\boldsymbol{\beta}_t = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)'$.

In the simulation we assume that all coordinates of vector $\tilde{X}$ are i.i.d. and follow the standard normal distribution $\mathcal{N}(0, 1)$. Then $U = \tilde{X}^T\tilde{\beta} \sim \mathcal{N}(0, ||\beta||_2^2)$. This means that for all $p$ the same results are obtained as in the algorithm only the knowledge of the distribution of $U$ is used. Let $F_{N(0,1)}(\cdot)$ denote distribution function of standard normal random variable and $F_{Cauchy(u,v)}(\cdot)$ distribution function of Cauchy distribution with location $u$ and scale $v$. In the case of incorrect model specification, the following response functions are considered:

$$\begin{aligned}
q_1(s) &= F_{N(0,1)}(s) \quad \text{(Probit model)},\\
q_2(s) &= \begin{cases} F_{N(0,1)}(s) & \text{for } F_{N(0,1)}(s) \in (0.1, 0.8)\\ 0.1 & \text{for } F_{N(0,1)}(s) \le 0.1\\ 0.8 & \text{for } F_{N(0,1)}(s) \ge 0.8, \end{cases}\\
q_3(s) &= \begin{cases} F_{N(0,1)}(s) & \text{for } F_{N(0,1)}(s) \in (0.2, 0.7)\\ 0.2 & \text{for } F_{N(0,1)}(s) \le 0.2\\ 0.7 & \text{for } F_{N(0,1)}(s) \ge 0.7, \end{cases}\\
q_4(s) &= \begin{cases} F_{N(0,1)}(s) & \text{for } |s| > 1\\ 0.5 + 0.5\cos[4\pi s]F_{N(0,1)}(s) & \text{for } |s| \le 1, \end{cases}\\
q_5(s) &= F_{Cauchy(0,1)}(s),\\
q_6(s) &= F_{Cauchy(0,2)}(s),
\end{aligned}$$

In Tables 1-2 values of $\beta_0^*$ and $\eta$ for models M1-M5, functions $q_1 - q_6$ and $q_L$ are given. Integrals were computed using Gauss-Hermite quadrature with 1000 nodes. No more than 7 iterations of the procedure were needed for convergence. Note that numerically calculated values of $\beta_0^*$ for $q_L, q_1, q_5$ and $q_6$ and all models considered are of order $10^{-16}$

or lower suggesting that values of $\beta_0^*$ are zero in these cases. This is indeed so due to symmetry of $X$ and the fact that these functions satisfy $q(x) = 1 - q(-x)$. We compare the results for $\eta$ with simulated values given in [19], reproduced here for convenience in Table 3. We observe that for all functions except $q_4$ (non-monotonic case) the results of both calculations are very close in terms of Mean Squared Error (given in the last row of Table 3), what suggest that the above Newton-Raphson procedure performs well for monotone $q$. We stress that values of $\eta$ different from 1 indicate misspecification and values significantly larger than 1 suggest that identification of an active set is easier when logistic model is fitted instead of the correct one. This surprising conclusion is indeed confirmed when Positive Selection rate (PSR) is considered as a measure of accuracy of detection of an active set (cf Figure 2 in [19]).
Note that Monte-Carlo calculation of $\eta$ performed in [19] was based on $10^6$ observations drawn from distribution of $(X, Y)$ whereas here we use a single run of the iterative procedure for its evaluation.

Table 1. Values of $\eta$ for models M1-M5 calculated by Newton-Raphson procedure.

| | $q_L$ | $q_1$ | $q_2$ | $q_3$ | $q_4$ | $q_5$ | $q_6$ |
|---|---|---|---|---|---|---|---|
| M1 | 1.0000 | 1.6041 | 1.6041 | 1.5977 | -0.1814 | 1.2462 | 0.6330 |
| M2 | 1.0000 | 1.7526 | 0.8655 | 0.5326 | 1.3211 | 0.8696 | 0.5244 |
| M3 | 1.0000 | 1.6836 | 1.3488 | 0.9634 | 0.8832 | 1.0504 | 0.5893 |
| M4 | 1.0000 | 1.6487 | 1.5300 | 1.2386 | 0.4861 | 1.1305 | 0.6107 |
| M5 | 1.0000 | 1.7503 | 0.8841 | 0.5461 | 1.3263 | 0.8772 | 0.5275 |

Table 2. Values of $\beta_0^*$ for models M1-M5 calculated by Newton-Raphson procedure.

| | $q_L$ | $q_1$ | $q_2$ | $q_3$ | $q_4$ | $q_5$ | $q_6$ |
|---|---|---|---|---|---|---|---|
| M1 | 4.39E-16 | 4.32E-16 | -6.04E-07 | -3.78E-04 | 4.25E-02 | -1.43E-17 | -1.91E-18 |
| M2 | -2.08E-16 | 4.41E-18 | -1.54E-01 | -1.66E-01 | -4.63E-02 | -2.05E-16 | -1.33E-16 |
| M3 | -1.20E-16 | -2.06E-16 | -5.85E-02 | -9.95E-02 | 1.38E-02 | 3.90E-16 | 4.16E-16 |
| M4 | 4.10E-16 | -1.41E-16 | -1.87E-02 | -5.62E-02 | 3.33E-04 | -8.56E-17 | -2.84E-17 |
| M5 | -2.23E-16 | 4.43E-17 | -1.51E-01 | -1.64E-01 | -6.16E-02 | -1.94E-16 | -1.30E-16 |

Table 3. Simulated values of $\hat{\eta}$ for considered models (reproduced from [19] (first 5 rows) together with MSEs between simulated and numercically calculated values given in the last row.

| | $q_L$ | $q_1$ | $q_2$ | $q_3$ | $q_4$ | $q_5$ | $q_6$ |
|---|---|---|---|---|---|---|---|
| M1 | 0.988 | 1.642 | 1.591 | 1.591 | 0.788 | 1.241 | 0.651 |
| M2 | 1.005 | 1.741 | 0.863 | 0.537 | 1.735 | 0.874 | 0.522 |
| M3 | 0.993 | 1.681 | 1.352 | 0.968 | 1.524 | 1.045 | 0.580 |
| M4 | 1.005 | 1.644 | 1.510 | 1.236 | 1.293 | 1.140 | 0.610 |
| M5 | 1.013 | 1.779 | 0.897 | 0.552 | 1.724 | 0.879 | 0.532 |
| MSE$\times 10^2$ | 0.008 | 0.049 | 0.016 | 0.003 | 46.600 | 0.003 | 0.009 |

### 5.2. *Omission of a valid predictor in a logistic regression model*

Our aim is now to check how omission of an active variable in logistic regression model affects other variables, in particular the inactive ones, which are fitted. Consider ten-dimensional random variable $\tilde{X} \sim \mathcal{N}(0, \Sigma)$,

$$\Sigma = \begin{bmatrix} I_2 & O \\ O & \Sigma_0 \end{bmatrix}, \Sigma_0 = [\rho^{|i-j|}]_{1 \le i,j \le 8}, |\rho| < 1$$

and logistic regression model

$$P(Y=1|X) = q_L(\cos\theta X_1 + \sin\theta X_2 + X_3). \tag{33}$$

Thus in this case $t \subseteq \{1,2,3\}$ (with equality holding when $\cos\theta\sin\theta \neq 0$), $X_3$ is always an active predictor and $X_4,\ldots,X_{10}$ are inactive. Consider now the case when $X_3$ is erroneously omitted from the fit of the logistic regression. The fact that $X_3$ is an active variable in the model is easily detectable by comparing residual deviances for the model $M_1$ containing all variables and $M_2$ containing all variables but $X_3$. Indeed, in the numerical experiments below the difference of deviances between these two models $dev_{M_1,M_2}$ is significant in 98.7% of the cases for $\rho = 0$ and in 90.9% of the cases for $\rho = 0.6$ at significance level $\alpha = 0.05$.
Denote by $X_{-3} = (1, X_1, X_2, X_4, \ldots, X_{10})^T$ and $\beta^*_{-3} = (\beta^*_0, \beta^*_1, \beta^*_2, \beta^*_4, \ldots, \beta^*_{10})^T$. Normal equations (12) for the model with $X_3$ omitted are

$$Eq_L(X^T_{-3}\beta^*_{-3})X_{-3} = Eq_L(\cos\theta X_1 + \sin\theta X_2 + X_3)X_{-3}.$$

Note that variables $X_3,\ldots,X_{10}$ are autoregressive $AR(1)$, thus $X_i = \rho^{i-3}X_3 + \sum_{j=4}^{i}\rho^{i-j}\varepsilon_j$ for $i \geq 4$, where $(\varepsilon_i)$ are iid $\mathcal{N}(0,1)$. Moreover, $X_1, X_2, X_3, \varepsilon_4, \ldots, \varepsilon_{10}$ are independent and hence using Stein's lemma we obtain for $i \geq 4$:

$$\begin{aligned} & \operatorname{Cov}(q_L(\cos\theta X_1 + \sin\theta X_2 + X_3), X_i) = \rho^{i-3}\operatorname{Cov}(q_L(\cos\theta X_1 + \sin\theta X_2 + X_3), X_3) \\ = \;& \rho^{i-3}Eq'_L(\cos\theta X_1 + \sin\theta X_2 + X_3). \end{aligned}$$

Again from Stein's lemma we have for $i \geq 4$:

$$\operatorname{Cov}(q_L(X^T_{-3}\beta^*_{-3}), X_i) = Eq'_L(X^T_{-3}\beta^*_{-3})\cdot\operatorname{Cov}(X^T_{-3}\beta^*_{-3}, X_i) = Eq'_L(X^T_{-3}\beta^*_{-3})\cdot\sum_{j=4}^{10}\beta^*_j\rho^{|j-i|}.$$

Thus from these equations and normal equations we get that:

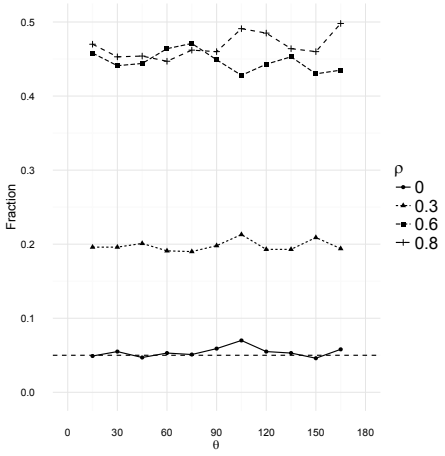
$$Eq'_L(X^T_{-3}\beta^*_{-3})\cdot\Sigma_0\cdot\begin{bmatrix}\beta^*_4\\ \vdots\\ \beta^*_{10}\end{bmatrix} = Eq'_L(\cos\theta X_1 + \sin\theta X_2 + X_3)\cdot\begin{bmatrix}\rho\\ \vdots\\ \rho^7\end{bmatrix}.$$

As matrix $\Sigma_0$ is invertible it is easy to check that the only solution to normal equation satisfies $\beta^*_5 = \ldots = \beta^*_{10} = 0$ and $\beta^*_4 = C\rho$, where:

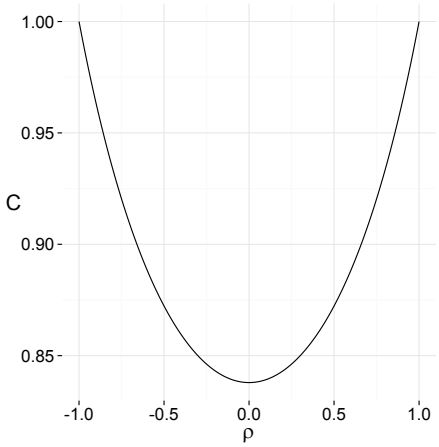
$$C = \frac{Eq'_L(\cos\theta X_1 + \sin\theta X_2 + X_3)}{Eq'_L(\beta^*_0 + \beta^*_1 X_1 + \beta^*_2 X_2 + \beta^*_4 X_4)}. \tag{34}$$

From symmetry of distribution of $X$, the fact that $q_L(s) = 1 - q_L(-s)$ and the uniqueness of the projection we have $\beta^*_0 = 0$. Moreover, it follows from Stein's lemma that $\beta^*_1 = C\cos\theta, \beta^*_2 = C\sin\theta$. This means that in our model $(\beta^*_1, \beta^*_2)^T = C(\beta_1, \beta_2)^T$ and $X_4$ takes over the role of omitted $X_3$. Note that $C$ does not depend on $\theta$, because $\cos\theta X_1 + \sin\theta X_2 = U \sim \mathcal{N}(0,1)$ and $\beta^*_1 X_1 + \beta^*_2 X_2 = CU$ and thus distribution of random variables appearing as arguments in (34), namely $U + X_3$ and $C(U + \rho X_4)$ do not depend on $\theta$.

Figure 2.




(a) Fraction of times inactive variable $X_4$ is significant after omission of valid variable versus $\theta$ for different values of $\rho$




(b) Constant $C$ versus correlation coefficient $\rho$ (see text)

Equation (34) can be thus restated as

$$C = \frac{Eq'_L(U + X_3)}{Eq'_L(C(U + \rho X_4)}.$$

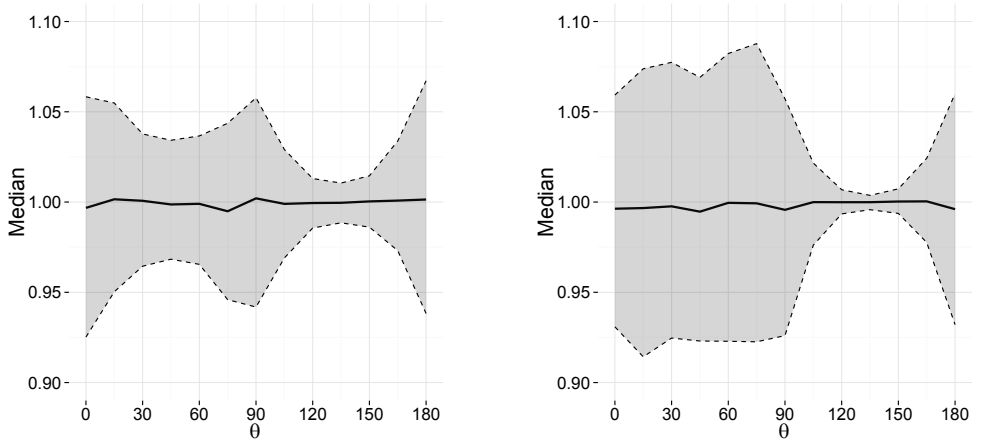
In the numerical study we generated $L_X = 10$ independent samples $\tilde{X}_1^{(j)}, \ldots, \tilde{X}_n^{(j)}$ $(j = 1, \ldots, L_X)$ with observations distributed as $\tilde{X}$ for $n = 100$ and the corresponding responses $Y_1^{(j,i)}, \ldots, Y_n^{(j,i)}$ $(i = 1, \ldots, L_Y$, where $L_Y = 100)$ were generated according to (33). Coefficients $\hat{\beta}^{(j,i)}$, where $j = 1, \ldots, L_X$, $i = 1, \ldots, L_Y$ are obtained from the logistic fit. In Figure 2(a) fraction of times when of $\hat{\beta}_4^{(j,i)}$ was significant according to Wald's test with significance level $\alpha = 0.05$ is plotted. From equation (34) we see that $C$ and thus $\beta_4^*$ does not on depend on $\theta$, what can be also inferred from the figure. It also indicates that the fraction of time $X_4$ is significant increases with $\rho$ for $\rho \in [0, 0.8]$. For $\rho = 0.9$ we obtain smaller fraction than for $\rho = 0.6$ due to approximate collinearity of $X_4$ and $X_5$ and resulting instability of $\hat{\beta}_4$. Panel (b) indicates that the dependence of $C$ on $\rho$ is approximately parabolic, although we were not able to justify it theoretically. As $C < 1$ this confirms a known dampening effect of omitting valid covariate on coefficient values of other true covariates.

### 5.3. *Connection between a linear filter and a logistic fit for normal distribution*

In the last example we check numerically that the equations (30) and (31) are approximately satisfied for two-dimensional predictors generated from $\mathcal{N}(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Figure 3.




(a) Median of $\alpha$ versus $\theta$ for $\rho = 0.3$ (see text) (b) Median of $\alpha$ versus $\theta$ for $\rho = 0.7$ (see text)

We generated $n = 100$ independent observations $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(n)}$ from this distribution and $Y_1, \ldots, Y_n$ such that $P(Y_i = 1|X_i) = q_L(\cos\theta X_1^{(i)} + \sin\theta X_2^{(i)})$ for $i = 1, \ldots, n$. We have fitted univariate linear models: $Y \sim X_1$, $Y \sim X_2$ and logistic model $Y \sim X_1 + X_2$. Obviously, it is easily detected that two first models are misspecified as assumed continuous response is fitted to binary outcome. Corresponding vector of slopes are denoted by $(\bar{\beta}^*_{1,lin}, \bar{\beta}^*_{2,lin})^T$, (linear model) and by $(\hat{\beta}^*_{1,log}, \hat{\beta}^*_{2,log})^T$ (logistic model). Then (cf. (31))

$$\hat{\beta} = \hat{\Sigma}^{-1}\,\mathrm{diag}(\hat{\Sigma}) \begin{bmatrix} \bar{\beta}^*_{1,lin} \\ \bar{\beta}^*_{2,lin} \end{bmatrix},$$

where $\hat{\Sigma}$ is an empirical covariance matrix.

In view of Remark 3.13 this vector should be approximately equal to $(\hat{\beta}^*_{1,log}, \hat{\beta}^*_{2,log})^T$, and in order to check this we calculated

$$\alpha = \frac{\hat{\beta}_1}{\hat{\beta}^*_{1,log}} \cdot \frac{\hat{\beta}^*_{2,log}}{\hat{\beta}_2},$$

which should be close to 1. Figure 3 depicting median of $\alpha$,versus $\theta$ based on $L = 2000$ repetitions shows that it is indeed so. We use the median here instead of the mean as for the considered small sample size ($n = 100$) distribution of $\alpha$ is skewed to the right. Lower and upper curves are the first and the third quartile $Q1(\alpha)$ and $Q3(\alpha)$, respectively.

## 6. Conclusion

In the paper we studied properties of Kullback-Leibler projection $\beta^*$ of the binary model on the logistic model and in particular the problem how the pertaining active set $t^*$ for the projection relates to the set of active predictors $t$. In Theorem 3.4 we proved that

KL projections are continuous with respect to distribution of predictors. We have shown that although in general the interplay between $t$ and $t^*$ can be arbitrary (Section 4), under some additional conditions such that of Corollary 3.2 we can assert that $t^* \subseteq t$. Moreover, in Section 3.2 we studied in detail the case when regressions of predictors are linear and show in this case that $t^* = t$ is ensured by the condition $\mathrm{Cov}(Y, X^T\beta) \neq 0$. We also studied related questions how removal of part of predictors influences KL projection (Proposition 3.9 ) and how KL projections on logistic and linear model compare (Proposition 3.11). In Section 5.2 we studied in detail projection on the logistic model in the case when misspecification rsults from an omission of an active predictor. We have also demonstrated that calculation of Ruud's proportionality constant $\eta$ by means of Newton-Raphson procedure for normal predictors is possible. This is potentially useful as values of $\eta$ significantly larger than 1 indicate that model misspecification can be beneficial for detection of the active set.
There are some interesting open questions which deserve closer scrutiny. Among others, it is unknown whether some relaxed form of Ruud's condition leads to an approximate proportionality of the true and projected vector of parameters. Moreover, other conditions than Ruud's condition on distribution of predictors and/or response function $q$ which lead to equality $t = t^*$ are unknown and worth investigating.

## 7. Appendix

Proof of Lemma 3.5.

*Proof.* By Theorem 3.4 we have $\beta^*(Z_m) \to \beta^*(Z)$ and moreover we know that $t^*(Z) = \{0, 1, \ldots, p\}$. Thus for all $i = 1, \ldots, p$ we have $\beta_i^*(Z) \neq 0$, and hence for sufficiently large $m$ we have $\beta_i^*(Z_m) \neq 0$. ■

Proof of Lemma 3.6.

*Proof.* Let us define $f(x) = q_L^{-1}(q(x))$ which by assumptions is a nonlinear function. Our goal is to define linearly nondegenerate random vector $Z = (Z_1, Z_2, \ldots, Z_p)^T$ such that

$$Z_1 + \ldots + Z_p = f(Z_1 + \ldots + Z_k). \tag{35}$$

Then it is obvious that with $\beta_i = I\{i \leq k\}\, i = 1, \ldots, p$ we will have $t(Z) = \{0, 1, \ldots, k\}$ and $t^*(Z) = \{0, 1, \ldots, p\}$. To this end let $\Omega = \{\omega_1, \ldots, \omega_{p+1}\} \subset R$ and $P(\{\omega_i\}) = 1/(p+1)$ for all $i = 1, \ldots, p+1$.
For $u_1, u_2 \in R$ to be specified later define

$$Z_i(\omega_j) = \begin{cases} 1 \Leftrightarrow 1 \leq i = j < p \\ 0 \Leftrightarrow (i \neq j \wedge 1 \leq i, j < p) \vee (j \in \{p, p+1\} \wedge 2 \leq i < p) \\ u_1 \Leftrightarrow i = 1 \wedge j = p \\ u_2 \Leftrightarrow i = 1 \wedge j = p+1 \\ f(1) - 1 \Leftrightarrow i = p \wedge 1 \leq j \leq k \\ f(0) - 1 \Leftrightarrow i = p \wedge k+1 \leq j \leq p-1 \\ f(u_1) - u_1 \Leftrightarrow i = p \wedge j = p \\ f(u_2) - u_2 \Leftrightarrow i = p \wedge j = p+1. \end{cases}$$

Then $Z = (Z_1, \ldots, Z_p)^T$ satisfies (35). Now we will choose $u_1, u_2$ such that $Z$ is linearly non-degenerate. From the definition of $Z$ this condition is equivalent to non-singularity

of the matrix: $A = [J_{p+1}|B]$, where $J_{p+1} = (1, \ldots, 1)^T \in R^{(p+1)\times 1}$ and $B = [Z_i(\omega_j)] \in R^{(p+1)\times p}$. We have

$$A = \begin{bmatrix} 1 & 1 & \ldots & 0 & 0 & \ldots & 0 & f(1)-1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \ldots & 1 & 0 & \ldots & 0 & f(1)-1 \\ 1 & 0 & \ldots & 0 & 1 & \ldots & 0 & f(0)-1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 & 0 & \ldots & 1 & f(0)-1 \\ 1 & u_1 & 0 & \ldots & \ldots & \ldots & 0 & f(u_1)-u_1 \\ 1 & u_2 & 0 & \ldots & \ldots & \ldots & 0 & f(u_2)-u_2 \end{bmatrix}.$$

It is seen that $|\det A| = |(1-u_2)(f(u_1)-u_1 f(1)) - (1-u_1)(f(u_2)-u_2 f(1))|$. From nonlinearity of $f$ there exists $u_2$ such that $f(u_2) \neq u_2 f(1)$. Obviously, $u_2 \neq 1$. Determinant $\det A$ is 0 if and only if for all $u_1 \in R$ :

$$f(u_1) = \frac{f(u_2)-f(1)}{u_2-1}u_1 + \frac{u_2 f(1)-f(u_2)}{u_2-1} =: \alpha u_1 + \beta.$$

From nonlinearity of $f$ the equality above does not hold for a certain $u_1 \neq 1$, otherwise we would have $\det A = 0$. If $u_1 \neq u_2$, this ends the proof. If not, from the continuity of $f(x) - \alpha x - \beta$ it follows that $f(u_2) \neq \alpha u_2 + \beta$ implies that there exists some $u_0 \neq u_2$ in the neighbourhood of $u_2$ such that $f(u_0) \neq \alpha u_0 + \beta$,. Taking $u_1 := u_0$ we have $\det A \neq 0$. This ends the proof. ■

*Acknowledgement* We are grateful to Piotr Pokarowski for fruitful discussions. The present form of Proposition 3.10 is due to him. Comments of two referees and Associate Editor which helped to improve the original manuscript are also gratefully acknowledged.

## References

[1] D. Brillinger. A generalized linear model with 'Gaussian' regressor variables. *Festschirift for Erich Lehmann*, pages 97–114, 1983.

[2] P. Brockwell and R. Davis. *Time Series: Theory and Methods*. Springer, New York, 1991.

[3] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, New York, 2012.

[4] P. Bühlmann and S. van de Geer. High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9:1449–1473, 2015.

[5] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

[6] R. Carroll and S. Pederson. On robustness in the logistic regression model. *Journal of the Royal Statistical Society B*, 55:693–706, 1993.

[7] L. Fahrmeir. Maximum likelihood estimation in misspecified generalized linear models. *Statistics*, 4:487–502, 1990.

[8] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics*, 1(13):342–368, 1985.

[9] J. Fan. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2005.

[10] J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, 38:35567–3604, 2010.

[11] M. Gail, W. Tan, and S. Piantadosi. Test for no treatment effect in randomized critical trials. *Biometrika*, 75:57–64, 1988.
[12] P. Hall and K. Li. On almost linearity of low dimensional projection from high dimensional data. *The Annals of Statistics*, 21(2):867–889, 1993.
[13] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, New York, 2015.
[14] N. Hjort and D. Pollard. Asymptotics for minimisers of convex processes. Unpublished manuscript, 1993. URL `http://www.stat.yale.edu/ pollard/Papers/convex.pdf`.
[15] J. Huang and C.H. Zhang. Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, 13:1839–1864, 2012.
[16] K. Li and N. Duan. Regression analysis under link violation. *The Annals of Statistics*, 17 (3):1009–1052, 1989.
[17] K. Li and N. Duan. Slicing regression: A link-free regression method. *The Annals of Statistics*, 19(2):505–530, 1991.
[18] J. Lv and J. Liu. Model selection principles in misspecified models. *Journal of Royal Statistical Society B*, 76:141–167, 2014.
[19] J. Mielniczuk and P. Teisseyre. What do we choose when we err ? Model selection and testing for misspecified logistic regression revisited. *Studies in Computational Intelligence*, 605:527–533, 2015.
[20] L. Robinson and N. Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 59:227–240, 1991.
[21] P. A. Ruud. Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica*, 51(1):225–228, 1983.
[22] C. Stein. Estimation of the mean of the multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151, 1981.
[23] S. van de Geer. *Estimation and Testing under Sparsity*. Springer, New York, 2016.
[24] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–84, 2004.