WARSAW UNIVERSITY OF TECHNOLOGY

Faculty of Mathematics and Information Science

PHD THESIS

mgr Mariusz Kubkowski

Misspecification of binary regression model: properties and inferential procedures

Thesis advisor

prof. dr hab. Jan Mielniczuk

WARSAW, 2019

Podziękowania

Pragnę serdecznie podziękować Panu prof. dr hab. Janowi Mielniczukowi za nieocenioną pomoc, życzliwość, cenne uwagi merytoryczne i wsparcie, na które zawsze mogłem liczyć.

Składam podziękowania Panu prof. dr hab. Jackowi Wesołowskiemu, dr Bartoszowi Kołodziejkowi oraz dr Wojciechowi Rejchelowi za okazaną pomoc i cenne sugestie, bez których ta praca nie byłaby kompletna.

Pragnę też podziękować mojej Rodzinie, a w szczególności: Mamie, ś.p. Tacie, Dziadkom oraz Siostrze, którzy zawsze we mnie wierzyli i mocno mnie wspierali.

Streszczenie

W poniższej rozprawie doktorskiej została przedstawiona problematyka złej specyfikacji modelu regresji binarnej. Pracę możemy podzielić zasadniczo na 4 części. W pierwszej części, którą stanowi Rozdział 1, został zawarty ogólny opis tego problemu oraz przykłady sytuacji, w których zła specyfikacja może wystąpić.

W drugiej części omówiono własności wektora współczynników teoretycznych $\boldsymbol{\beta}^*$ w dopasowanym modelu - wyniki zawarte w tej części stanowią uogólnienie wyników zawartych w pracach Kubkowski, Mielniczuk (2017) (Rozdział 2) oraz Kubkowski, Mielniczuk (2018) (Rozdział 3) do przypadku wypukłej funkcji straty. W Rozdziale 2 zbadano własności nośnika s^* wektora współczynników teoretycznych w dopasowanym modelu w przypadku spełnienia warunku liniowych regresji i w przypadku niespełnienia tego warunku. W Rozdziale 3 jest rozważany ponadto addytywny model binarny.

Trzecia część, składająca się z Rozdziałów 4 i 5, skupia się na estymacji wektora $\boldsymbol{\beta}^*$ oraz zbioru s^* dla losowych predyktorów subgaussowskich (także w przypadku, gdy liczba predyktorów jest większa od liczby obserwacji). W Rozdziale 4 pokazano wyniki dotyczące metody Lasso oparte o idee zawarte w pracach Fan i in. (2014a) oraz Bühlmann, van de Geer (2011). W Rozdziale 5 omówiono minimalizację Uogólnionego Kryterium Informacyjnego (GIC) w pewnej rodzinie $\mathcal{M}$, do której należy s^*. W Rozdziale 5 przedstawiono także procedurę dwustopniową SS (Screening - Selection) służącą do znajdowania estymatora s^*, która opiera się w swoim działaniu o metodę Lasso (pierwszy etap) i minimalizację GIC (drugi etap). W Rozdziale 5 zaprezentowano także rezultaty teoretyczne dotyczące jej działania.

Czwarta część (Rozdział 6) zawiera opisy i analizę eksperymentów numerycznych, w których zbadano procedury będące modyfikacjami procedury SS dla próby losowej oraz zaprezentowano procedurę numerycznego przybliżenia $\boldsymbol{\beta}^*$ i sprawdzono numerycznie jej działanie.

Słowa kluczowe: zła specyfikacja, binarny model regresyjny, regresja logistyczna, Lasso, Uogólnione Kryterium Informacyjne, zbiory aktywnych predyktorów, selekcja zmiennych, regresja wysoko-wymiarowa.

Abstract

In this doctoral dissertation problem of misspecification of binary regression model is discussed. This dissertation consists of four parts. In the first part, consisting of Chapter 1, general description of this problem and examples of situations, where misspecification occurs, are given.

In the second part, we discuss properties of vector of theoretical coefficients $\boldsymbol{\beta}^*$ in fitted model. Results presented in this part generalize results contained in Kubkowski, Mielniczuk (2017) (Chapter 2) and Kubkowski, Mielniczuk (2018) (Chapter 3) to the case of convex loss function. In Chapter 2 we study properties of support s^* of $\boldsymbol{\beta}^*$ in fitted model in the case when linear regressions condition is satisfied and in the case when this condition is not satisfied. We consider additionally additive binary model in Chapter 3.

In third part, consisting of Chapters 4 and 5, we focus on estimation of vector $\boldsymbol{\beta}^*$ and set s^* for random subgaussian predictors (also in the case when number of predictors is greater than number of observations). In Chapter 4 several novel results concerning Lasso are shown. The results are based on ideas contained in papers of Fan et al (2014a) and Bühlmann, van de Geer (2011). In Chapter 5 minimization of Generalized Information Criterion over family $\mathcal{M}$ (to which s^* belongs) is discussed. In Chapter 5 two-stage SS (Screening - Selection) procedure of finding estimator of s^* is presented and its selection consistency is discussed. The procedure consists of screening based on Lasso in the first stage and GIC minimization in the second stage. In Chapter 5 theoretical results concerning SS procedure are presented.

Fourth part (Chapter 6) contains description and analysis of numerical experiments, in which we study properties of procedures which are modifications of SS procedure. We also present in this chapter procedure approximating $\boldsymbol{\beta}^*$ numerically and we check its performance.

Key words: misspecification, binary regression model, logistic regression, Lasso, Generalized Information Criterion, sets of active predictors, variable selection, high-dimensional regression.

Contents

Notation and conventions

- Random variables are denoted by big letters, e.g. $X_1, X_2, Y, Z, \ldots$,
- vectors are additionally denoted in bold font, e.g. $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}, \mathbf{Z}, \ldots$,
- observations of random variables are denoted in small font, e.g. $x_1, x_2, y, z, \ldots$,
- observations of random vectors are additionally denoted in bold font, e.g. $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, \mathbf{z}, \ldots$,
- $\mathbf{0}_p$ - $p \times 1$ vector of zeros,
- $\mathbf{O}_{p\times m}$ - $p \times m$ matrix of zeros,
- $\mathbb{N}$ - set of natural numbers (including 0), $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$,
- Df - gradient of function f,
- $\Phi(x) = \int\limits_{-\infty}^{x} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$ for $x \in \mathbb{R}$ - cdf of $\mathcal{N}(0,1)$ distribution,
- $\phi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ for $x \in \mathbb{R}$ - pdf of $\mathcal{N}(0,1)$ distribution,
- $\boldsymbol{\beta}$ - vector of true values of parameters,
- $\boldsymbol{\beta}^*$ - projection vector,
- $\hat{\boldsymbol{\beta}}_L$ - Lasso estimator of $\boldsymbol{\beta}^*$,
- $\hat{\boldsymbol{\beta}}(w)$ - ML estimator calculated on model $w \subseteq \{1, \ldots, p\}$,
- $\tilde{\mathbf{v}}$ - vector $\mathbf{v}$ with omitted first coordinate,
- $\mathbf{v}_\pi = (v_{j_1}, \ldots, v_{j_k})^T$ -subvector of $\mathbf{v} \in \mathbb{R}^p$ and $\pi = \{j_1, \ldots, j_k\} \subseteq \{1, \ldots, p\}$, $(\mathbf{v}_\emptyset = 0)$,
- $\mathbb{X}_\pi = \mathbb{X}^{(j_1,\ldots,j_k)}$ - submatrix of $\mathbb{X} \in \mathbb{R}^{n\times p}$ with columns indexed by elements of $\pi = \{j_1, \ldots, j_k\} \subseteq \{1, \ldots, p\}$, $(\mathbb{X}_\emptyset = \mathbf{0}_n)$,
- s - set of true active predictors,
- s^* - set of active predictors corresponding to $\boldsymbol{\beta}^*$,
- $q, q^{(n)}$ -response function,
- $q_L(x) = (1 + \exp(-x))^{-1}$, $x \in \mathbb{R}$ - logistic function,
- $I(A)$ - characteristic function of set A,
- $a_+ = aI(a > 0)$ and $a_- = aI(a < 0)$ for $a \in \mathbb{R}$,
- $|w|$ - cardinality of set w.

Chapter 1

Introduction

Let $n \in \mathbb{N}_+, p_n \in \mathbb{N}$, $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \sim \mathbb{P} = \mathbb{P}_n$ be i.i.d. random variables, $(\mathbf{X}, Y) \in \mathbb{R}^{p_n+1} \times \{0, 1\}$. We consider a general binary model such that a conditional distribution of Y given $\mathbf{X}$ is given by

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q^{(n)}(\mathbf{x}), \tag{1.1}$$

where $\mathbf{X} = (X_0, X_1, \ldots, X_{p_n})^T$ is a column vector of predictors, $X_0 \equiv 1$ and $q^{(n)}\colon\ \mathbb{R} \to [0, 1]$ is a certain unknown response function. Note that variable X_0 is associated with intercept in the regression model.

Let Θ be a set of possible parameter values,

$$l\colon\ \Theta \times \mathbb{R}^{p_n+1} \times \{0, 1\} \to \mathbb{R}$$

be a loss function with $R : \Theta \to \mathbb{R}$ associated risk function given by

$$R(\mathbf{b}) = \mathbb{E}l(\mathbf{b}, \mathbf{X}, Y). \tag{1.2}$$

Object of main interest here is the minimizer of the risk:

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}^*_n = \arg\min_{\mathbf{b}\in\Theta} R(\mathbf{b}). \tag{1.3}$$

Consider set of active predictors corresponding to $\boldsymbol{\beta}^*$:

$$s^* = \{i \in \{1, \ldots, p_n\}\colon\ \exists \mathbf{x} \in \mathbb{R}^{p_n+1}, x'_i \in \mathbb{R}, y \in \{0, 1\}\colon\ l(\boldsymbol{\beta}^*, \mathbf{x}, y) \neq l(\boldsymbol{\beta}^*, \mathbf{x}'_{(i)}, y)\}, \tag{1.4}$$

where $\mathbf{x} = (1, x_1, \ldots, x_i, \ldots, x_{p_n})^T$ and $\mathbf{x}'_{(i)} = (1, x_1, \ldots, x'_i, \ldots, x_{p_n})^T$ is vector $\mathbf{x}$ with replaced $(i + 1)$-th coordinate corresponding to x_i by x'_i. To discuss the properties of s^* we have to assume that $\boldsymbol{\beta}^*$ exists and is uniquely defined. Conditions for this are given in Appendix A.1 for losses of the form $l(\mathbf{b}, \mathbf{x}, y) = \rho(\mathbf{b}^T\mathbf{x}, y)$.

Analogously, we define the following set of true active predictors:

$$s = \{i \in \{1, \ldots, p_n\}\colon\ \exists \mathbf{x} \in \mathbb{R}^{p_n+1},\ x'_i \in \mathbb{R}\colon\ q^{(n)}(\mathbf{x}) \neq q^{(n)}(\mathbf{x}'_{(i)})\}. \tag{1.5}$$

We now discuss the most important case of the above situation, when the loss function is associated with a fitted model. Namely, assume that we want to find the projection of the model (1.1) on the family of parametric models $\{\pi(\mathbf{x}, \mathbf{b})\colon\ \mathbf{b} \in \Theta\}$ (in this case $\boldsymbol{\beta}^*$ is a parameter corresponding to the projection) characterized by equation:

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \pi(\mathbf{x}, \mathbf{b}),$$

where $\pi\colon\ \mathbb{R}^{p_n+1} \times \Theta \to [0,1]$. We note that the last equality can be written as:

$$\mathbb{P}(Y = y|\mathbf{X} = \mathbf{x}) = \pi(\mathbf{x}, \mathbf{b})^y (1 - \pi(\mathbf{x}, \mathbf{b}))^{1-y}.$$

If for some $\boldsymbol{\beta} \in \Theta$ $\pi(\mathbf{x}, \boldsymbol{\beta}) = q^{(n)}(\mathbf{x})$ $\mathbb{P}_{\mathbf{X}}$ - a.e. and associated loss function is given as

$$-\log \mathcal{L}(\mathbf{b}, \mathbf{x}, y) = -\ln \mathbb{P}_{\mathbf{b}}(Y = y|\mathbf{X} = \mathbf{x}) = -y \ln(\pi(\mathbf{x}, \mathbf{b})) - (1 - y) \ln(1 - \pi(\mathbf{x}, \mathbf{b}))$$

then the model is well specified. More specifically, we have:

Definition 1.1. *We call binary model with loss function l well specified (in a general sense) with respect to the family of parametric binary models $\{\mathbb{P}_{\boldsymbol{b}}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x}, \boldsymbol{b})\}$, where $\pi\colon\ \mathbb{R}^{p_n+1} \times \Theta \to [0,1]$ if it satisfies the following two conditions:*

1. *For all $\boldsymbol{x} \in \mathbb{R}^{p_n+1}, \boldsymbol{b} \in \Theta, y \in \{0,1\}$ we have:*
$$l(\boldsymbol{b}, \boldsymbol{x}, y) = -y \log(\pi(\boldsymbol{x}, \boldsymbol{b})) - (1 - y) \log(1 - \pi(\boldsymbol{x}, \boldsymbol{b})).$$
2. *There exists vector $\boldsymbol{\beta} \in \Theta$ that for all $\boldsymbol{x} \in \mathbb{R}^{p_n+1}$: $q^{(n)}(\boldsymbol{x}) = \pi(\boldsymbol{x}, \boldsymbol{\beta})$.*

We say that binary model with loss function l is misspecified (in a general sense) with respect to the family of parametric binary models $\{\mathbb{P}_{\mathbf{b}}(Y = 1|\mathbf{X} = \mathbf{x}) = \pi(\mathbf{x}, \mathbf{b})\}$, if it is not well specified with respect to this family.

We note in particular that if $q(\mathbf{x})$ is itself a member of parametric family $\{\pi(\mathbf{x}, \mathbf{b}) : \mathbf{b} \in \Theta\}$, i.e. $q(\mathbf{x}) = \pi(\mathbf{x}, \boldsymbol{\beta})$ for some $\boldsymbol{\beta} \in \Theta$ and equality $l(\mathbf{b}, \mathbf{x}, y) = -\log \mathcal{L}(\mathbf{b}, \mathbf{x}, y)$ does not hold for all $\mathbf{b} \in \Theta$, then binary model is misspecified (in a general sense) with respect to parametric family $\{\pi(\mathbf{x}, \mathbf{b}) : \mathbf{b} \in \Theta\}$. In this case we simply call binary model misspecified. Thus the model corresponding to data generating mechanism is misspecified or well specified. We note that in Kubkowski and Mielniczuk (2017) different terminology was used. We give examples of misspecified models in Section 1.2.

In addition to the general model in (1.1) we consider two specific setups in this dissertation:

- semiparametric setup ($\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p_n})^T \in \mathbb{R}^{p_n+1}$, $q^{(n)}\colon\ \mathbb{R} \to \mathbb{R}$) - see Chapter 2:
$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q^{(n)}(\mathbf{x}^T \boldsymbol{\beta}), \tag{1.6}$$
(abusing notation slightly we will denote by $q^{(n)}$ a function satisfying (1.1) or (1.6)).
- generalized semiparametric setup ($k \in \mathbb{N}$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k \in \mathbb{R}^{p_n+1}$, $q^{(n)}\colon\ \mathbb{R}^k \to \mathbb{R}$) - see Chapter 3:
$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q^{(n)}(\mathbf{x}^T \boldsymbol{\beta}_1, \ldots, \mathbf{x}^T \boldsymbol{\beta}_k). \tag{1.7}$$

We note that semiparametric setup is equivalent in the binary case to the model often considered in literature (see e.g. Li and Duan (1989)): $Y = g(\boldsymbol{\beta}^T \mathbf{X}, \varepsilon)$ for some function g and ε independent of $\mathbf{X}$, what is shown in the Remark 2.8. We note that the construction discussed in the proof can be easily generalized to the case of generalized semiparametric setup.

In this thesis we consider regression type loss of the form $l(\mathbf{b}, \mathbf{x}, y) = \rho(\mathbf{b}^T\mathbf{x}, y)$, where $\mathbf{b} \in \Theta = \mathbb{R}^{p_n+1}, \rho\colon\ \mathbb{R} \times \{0,1\} \to \mathbb{R}$. In case of such loss we observe that

$$s^* = \{i \in \{1, \ldots, p_n\}\colon\ \beta_i^* \neq 0\}$$

(when ρ is not degenerate in the sense that $\rho(c, y) \not\equiv \rho(y)$).

For semiparametric and generalized semiparametric setup we note also that we have:

$$s = \{i \in \{1, \ldots, p_n\}\colon\ \beta_i \neq 0\},$$

$$s = \{i \in \{1, \ldots, p_n\}\colon\ \exists j \in \{1, \ldots, k\}\colon\ \beta_{ji} \neq 0\},$$

respectively (assuming $q^{(n)}$ is nondegenerate function in similar sense as ρ).

Remark 1.2. *In this dissertation we consider also model without intercept, when* $\Theta = \mathbb{R}^{p_n+1}$. *In this case we define:*

$$\boldsymbol{\beta}^* = \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^{p_n}} R((0, \boldsymbol{b}^T)^T). \tag{1.8}$$

Proofs of all theorems will be given for the model with intercept (where $\boldsymbol{\beta}^*$ *is given by Equation (1.3)) unless it is specified differently.*

One of the main problems considered in this work is the interplay between sets s and s^*. We show that set s^* is a subset of set s (or even equal to it) in the semiparametric setup under linear regressions condition (see Chapter 2). Moreover, we provide examples that when linear regressions assumption is violated, the relation between sets s and s^* can be arbitrary. The results for these relations given here are mainly for semiparametric setup.

Next important problem considered here is the relation between vectors $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$ in semiparametric setup (or between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ in generalized semiparametric setup). Under linear regressions condition it turns out that $\boldsymbol{\beta}^*$ is proportional to $\boldsymbol{\beta}$ (or a linear combination of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$) - see Chapter 3. The last question, which we will try to answer is how to select the set which approximates s^* when we are given i.i.d. random sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ specified above. We introduce an appropriate procedure, establish its properties for subgaussian predictors and perform numerical experiments to check its effectiveness - see Chapters 4-6. Moreover, we show in numerical experiments (Chapter 6) that when the linear regressions condition is satisfied, set s^* can be selected correctly with high probability even for loss functions which are not associated with any particular model (for example Huber loss).

Properties of inferential procedures under misspecification when family of logistic models $\{q_L(\mathbf{x}^T\mathbf{b})\}$ for $\mathbf{b} \in \mathbb{R}^{p+1}$ is fitted ($q_L(x) = (1 + e^{-x})^{-1}$ for $x \in \mathbb{R}$), serves as a main example in the dissertation.

1.1. Basic loss functions

In this chapter we give definitions of several loss functions, which are usually considered in binary misspecification problem.

Logistic loss is the main loss function of our interest corresponding to logistic regression fit:

$$l_{log}(\mathbf{b}, \mathbf{x}, y) = -y\mathbf{x}^T\mathbf{b} + \ln\Big(1 + \exp\Big(\mathbf{x}^T\mathbf{b}\Big)\Big) = \rho_{log}(\mathbf{x}^T\mathbf{b}, y), \tag{1.9}$$

where $\rho_{log}(b, y) = -yb + \ln\Big(1 + e^b\Big)$. This loss equals $-\log \mathbb{P}(Y = y|\mathbf{X} = \mathbf{x})$ in the logistic regression model. We note that $\rho_{log}(\cdot, y)$ is non-negative, differentiable, strictly convex function for all y.

Probit loss is a loss related to probit regression:

$$l_{prob}(\mathbf{b}, \mathbf{x}, y) = -y \ln\Big(\Phi(\mathbf{x}^T\mathbf{b})\Big) - (1 - y) \ln\Big(1 - \Phi(\mathbf{x}^T\mathbf{b})\Big) = \rho_{prob}(\mathbf{x}^T\mathbf{b}, y), \tag{1.10}$$

where $\rho_{prob}(b, y) = -y \ln(\Phi(b)) - (1 - y) \ln(1 - \Phi(b))$ and:

$$\Phi(b) = \int\limits_{-\infty}^{b} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx.$$

We note that $\rho_{prob}(\cdot, y)$ is non-negative, differentiable, strictly convex function for all y (for the proof of strict convexity see Remark A.14).

Quadratic (or squared) loss is a loss related to the linear regression:

$$l_{lin}(\mathbf{b}, \mathbf{x}, y) = \frac{1}{2}(y - \mathbf{x}^T\mathbf{b})^2 = \rho_{lin}(\mathbf{x}^T\mathbf{b}, y), \tag{1.11}$$

where

$$\rho_{lin}(b, y) = \frac{1}{2}(y - b)^2.$$

It turns out that for quadratic loss we can give explicit formula for $\boldsymbol{\beta}^*$ (see (2.16)). We also note that $\rho_{lin}(\cdot, y)$ is non-negative, differentiable, strictly convex function for all y.

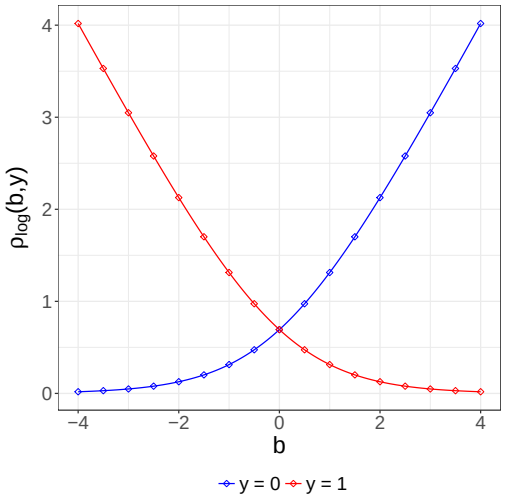
Huber loss is a loss related to Huber regression:

$$l_H^\delta(\mathbf{b}, \mathbf{x}, y) = \begin{cases} \frac{(y - \mathbf{x}^T\mathbf{b})^2}{2\delta} & |y - \mathbf{x}^T\mathbf{b}| \le \delta \\ |y - \mathbf{x}^T\mathbf{b}| - \frac{1}{2}\delta & |y - \mathbf{x}^T\mathbf{b}| > \delta \end{cases} = \rho_H^\delta(\mathbf{x}^T\mathbf{b}, y), \tag{1.12}$$

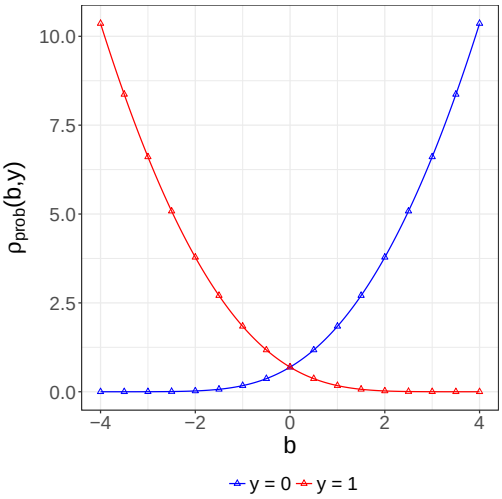
where $\delta > 0$ and

$$\rho_H^\delta(b, y) = \begin{cases} \frac{(y-b)^2}{2\delta} & |y - b| \le \delta \\ |y - b| - \frac{1}{2}\delta & |y - b| > \delta \end{cases}.$$

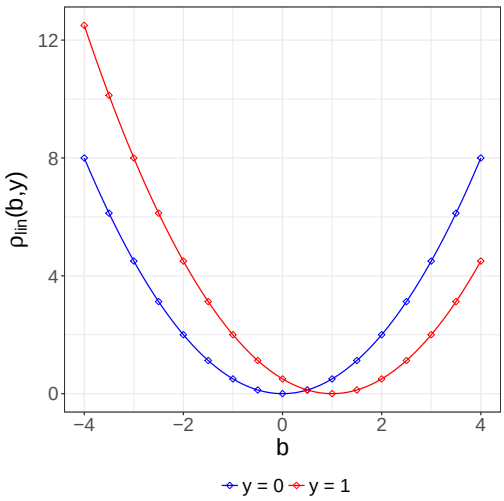
We note that $\rho_H^\delta(\cdot, y)$ is differentiable, Lipschitz, convex (but not strictly convex) function for all y.



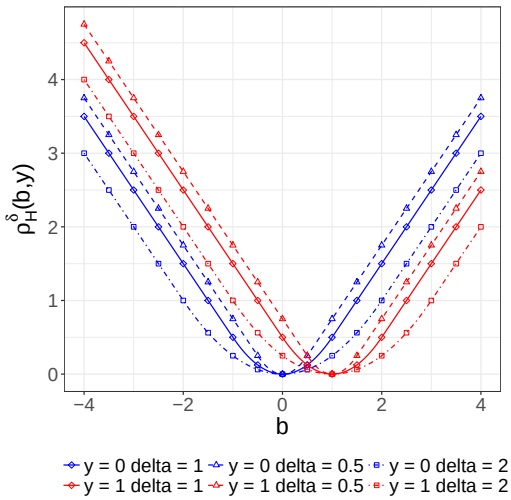
(a) Logistic loss



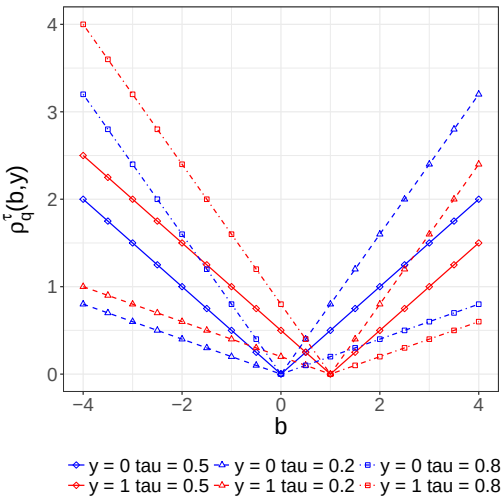
(b) Probit loss



(c) Quadratic loss



(d) Huber loss



(e) Quantile loss

Figure 1.1: Functions ρ for different losses.

Quantile loss is related to quantile regression:

$$l_q^\tau(\mathbf{b}, \mathbf{x}, y) = (y - \mathbf{x}^T\mathbf{b})(\tau - I(y - \mathbf{x}^T\mathbf{b} < 0)) = \rho_q^\tau(\mathbf{x}^T\mathbf{b}, y), \tag{1.13}$$

where $\tau \in (0, 1)$ and

$$\rho_q^\tau(b, y) = (y - b)(\tau - I(y - b < 0)).$$

Note that for $\tau = 1/2$ we have $l_q^\tau(\mathbf{b}, \mathbf{x}, y) = |y - \mathbf{x}^T\mathbf{b}|/2$. We observe that $\rho_q^\tau(\cdot, y)$ is Lipschitz, convex function for all y. However, $\rho_q^\tau(\cdot, y)$ is not differentiable.

1.2. Examples of misspecified models

In this section we provide examples of well specified and misspecified models in order to provide better understanding of the problem presented in this dissertation.

Example 1.3. *Let $X \sim \mathcal{N}(0, 1)$, $\Theta = \mathbb{R}^2$,*

$$\mathbb{P}(Y = 1|X = x) = \Phi(x) = \int_{-\infty}^{x} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt.$$

Observe that conditional distribution of Y given X corresponds to the probit model. Now we consider the logistic loss function. Then the model is misspecified, as for all $\boldsymbol{b} = (b_0, b_1)^T \in \Theta$ as we have $\Phi(x) \not\equiv q_L(b_0 + b_1 x)$.

Example 1.4. *Let X_1, X_2 be independent random variables, where $X_1 \sim \mathcal{N}(0, 1), X_2 \sim Bern(0.5)$ and let*

$$\mathbb{P}(Y = 1|X_1 = x_1, X_2 = x_2) = q_1(x_1, x_2) = \frac{1}{1 + e^{-x_1 - x_2}}.$$

We consider logistic loss function. Then the model with predictors X_1 and X_2 is well specified, as for

$$\boldsymbol{b} = (b_0, b_1, b_2)^T = (0, 1, 1)^T$$

we have

$$q_1(x_1, x_2) \equiv q_L(b_0 + b_1 x_1 + b_2 x_2).$$

However, when we omit variable X_2, then we obtain:

$$\begin{aligned}\mathbb{P}(Y = 1|X_1 = x_1) &= \mathbb{P}(Y = 1|X_1 = x_1, X_2 = 1)\mathbb{P}(X_2 = 1|X_1 = x_1) \\ &\quad + \mathbb{P}(Y = 1|X_1 = x_1, X_2 = -1)\mathbb{P}(X_2 = -1|X_1 = x_1) \\ &= \frac{1}{2} \cdot \frac{1}{1 + e^{-x_1 - 1}} + \frac{1}{2} \cdot \frac{1}{1 + e^{-x_1 + 1}} = q_2(x_1).\end{aligned}$$

This means that model with only predictor X_1 is misspecified, as for all $\boldsymbol{b} = (b_0, b_1)^T \in \mathbb{R}^2$ we have $q_2(x_1) \not\equiv q_L(b_0 + b_1 x_1)$.

Note that this example is a special case of Example 1.6 below.

Example 1.5. *Let X_1, X_2 be independent non-degenerate random variables and assume $X_1 \leq 0$,*

$$\mathbb{P}(X_2 = 1) = \mathbb{P}(X_2 = -1) = \frac{1}{2},$$

$$\mathbb{P}(Y = 1|X_1 = x_1, X_2 = x_2) = q(x_1, x_2) = \frac{2}{1 + e^{-x_1}} I(x_2 = 1).$$

We consider logistic loss function. Then model containing X_1 and X_2 as predictors is misspecified whereas model with only X_1 as predictor will be well specified as we have:

$$\begin{aligned}\mathbb{P}(Y = 1|X_1 = x_1) &= \mathbb{P}(Y = 1|X_1 = x_1, X_2 = 1)\mathbb{P}(X_2 = 1|X_1 = x_1) \\ &+ \mathbb{P}(Y = 1|X_1 = x_1, X_2 = -1)\mathbb{P}(X_2 = -1|X_1 = x_1) = \frac{1}{1 + e^{-x_1}}.\end{aligned}$$

Example 1.6. *Let $T \in \{1, \ldots, k\}$ and $\boldsymbol{X} \in \mathbb{R}^p$ be independent random variables and let for $i \in \{1, \ldots, k\}$ and $\boldsymbol{\beta}_{(i)} \in \mathbb{R}^p$:*

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}, T = i) = q_i(\boldsymbol{\beta}_{(i)}^T \boldsymbol{x}), \quad \mathbb{P}(T = i) = p_i \in (0, 1),$$

where $\sum_{i=1}^k p_i = 1$. We consider the following additive model:

$$\begin{aligned}\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) &= \sum_{i=1}^k \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}, T = i)\mathbb{P}(T = i|\boldsymbol{X} = \boldsymbol{x}) \\ &= \sum_{i=1}^k q_i(\boldsymbol{\beta}_{(i)}^T \boldsymbol{x})\mathbb{P}(T = i) = \sum_{i=1}^k p_i q_i(\boldsymbol{\beta}_{(i)}^T \boldsymbol{x}).\end{aligned}$$

We consider logistic loss function. We show for specific k,q_i,$\boldsymbol{\beta}_{(i)}$ and p_i that the introduced model is misspecified, if there exists $\boldsymbol{x}_0 \in \mathbb{R}^p$ such that $\boldsymbol{\beta}^T \boldsymbol{x}_0 > 0$ and $\operatorname{supp} \boldsymbol{X} \supseteq \{c\boldsymbol{x}_0 \in \mathbb{R}^p\colon\ c \in \mathbb{R}\}$. Let e.g. $k = 2$, $p_1 = 1 - p$, $p_2 = p$, $p \in (0, 1) \setminus \{0.5\}$, $\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}$, $\boldsymbol{\beta}_{(2)} = -\boldsymbol{\beta}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, $q_1 = q_2 = q_L$. Then:

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = (1 - p)q_L(\boldsymbol{\beta}^T \boldsymbol{x}) + pq_L(-\boldsymbol{\beta}^T \boldsymbol{x}). \tag{1.14}$$

We have additionally

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}, T = 1) = q_L(\boldsymbol{\beta}^T \boldsymbol{x}), \quad \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}, T = 2) = q_L(-\boldsymbol{\beta}^T \boldsymbol{x}).$$

This means that when $T = 2$ then the mislabelling of Y class (from 0 to 1 and vice versa) occurs. The probability of this event is $\mathbb{P}(T = 2) = p$. Now, we observe that in view of assumed condition:

$$\lim_{c \to +\infty} \left(pq_L(\boldsymbol{\beta}^T(\boldsymbol{x}_0 c)) + (1 - p)q_L(-\boldsymbol{\beta}^T(\boldsymbol{x}_0 c)) \right) = p \notin \left\{ 0, \frac{1}{2}, 1 \right\}.$$

However, we have for any $\boldsymbol{\gamma} \in \mathbb{R}^p$:

$$\lim_{c \to +\infty} q_L(\boldsymbol{\gamma}^T(\boldsymbol{x}_0 c)) = \begin{cases} 1, & \text{if } \boldsymbol{\gamma}^T \boldsymbol{x}_0 > 0 \\ \frac{1}{2}, & \text{if } \boldsymbol{\gamma}^T \boldsymbol{x}_0 = 0 \\ 0, & \text{if } \boldsymbol{\gamma}^T \boldsymbol{x}_0 < 0 \end{cases}.$$

Therefore we have shown that the model (1.14) with logistic loss is misspecified.

Chapter 2

Properties of the projection in the semiparametric model

In this chapter we consider semiparametric binary model:

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q^{(n)}(\mathbf{x}^T\boldsymbol{\beta}), \tag{2.1}$$

where $\mathbf{x}, \boldsymbol{\beta} \in \Theta = \mathbb{R}^{p_n+1}$. To simplify the considerations, we assume that $q^{(n)} = q\colon\ \mathbb{R} \to [0,1]$, $p_n = p \in \mathbb{N}$, $\Theta = \mathbb{R}^{p+1}$. Because our focus will be on non-constant predictors and not on the intercept, we introduce the following notation: $\mathbf{X} = (X_0, \tilde{\mathbf{X}}^T)^T$, $\tilde{\mathbf{X}} = (X_1, \ldots, X_p)^T$, $X_0 \equiv 1$, $\tilde{\mathbf{b}} = (b_1, \ldots, b_p)^T$, $\mathbf{b} = (b_0, \tilde{\mathbf{b}}^T)^T$ (and we define $\tilde{\boldsymbol{\beta}}$ analogously as $\tilde{\mathbf{b}}$). If the appropriate moments of $\tilde{\mathbf{X}}$ are finite, we will write $\mathbb{E}\tilde{\mathbf{X}} = \boldsymbol{\mu}$, $\operatorname{Var}\tilde{\mathbf{X}} = \boldsymbol{\Sigma}$.

One of the most important assumptions considered in this chapter is linear regressions condition:

$$(\mathrm{LRC}(\mathbf{b}))\ \ \exists \mathbf{h}_0 = \mathbf{h}_0(\tilde{\mathbf{b}}) \in \mathbb{R}^p, \mathbf{h}_1 = \mathbf{h}_1(\tilde{\mathbf{b}}) \in \mathbb{R}^p\colon\ \mathbb{E}(\tilde{\mathbf{X}}|\tilde{\mathbf{b}}^T\tilde{\mathbf{X}}) = \mathbf{h}_0 + \mathbf{h}_1\tilde{\mathbf{b}}^T\tilde{\mathbf{X}},$$

Condition LRC($\mathbf{b}$) is satisfied for every $\mathbf{b} \in \mathbb{R}^{p+1}$ with $\tilde{\mathbf{b}} \neq \mathbf{0}_p$ by elliptically contoured distributions of $\tilde{\mathbf{X}}$ having finite second moment (see Section A.2 in Appendix). Moreover, if the condition LRC($\mathbf{b}$) is satisfied for every $\mathbf{b} \in \mathbb{R}^{p+1}$, then $\tilde{\mathbf{X}}$ follows elliptically contoured distribution (see Theorem A.23 in Appendix). Note that normal distribution belongs to the family of elliptically contoured distributions (see Remark A.17 in Appendix).

We will show in the Theorem 2.6 that the condition LRC($\boldsymbol{\beta}$) is essential in the proof of equality $\tilde{\boldsymbol{\beta}}^* = \eta\tilde{\boldsymbol{\beta}}$ for some $\eta \in \mathbb{R}$. Condition LRC($\boldsymbol{\beta}^*$) in the case of logistic (or quadratic) loss below allows us to represent $\boldsymbol{\beta}^*$ as some function of $\boldsymbol{\beta}$ even in the situation when some relevant predictors are omitted in the fitted logistic (or linear) model - see Sections 2.2 and 2.3.

According to a discussion in Chapter 1 model (2.1) is misspecified when associated loss function is not equal to minus log-likelihood of this model. Such a case will be investigated in Chapter 2 for a general loss defined in (2.2) as well as for specific cases when l is logistic, probit and quadratic loss.

In this chapter we assume throughout that loss function is of the form:

$$l(\mathbf{b}, \mathbf{x}, y) = \rho(\mathbf{b}^T\mathbf{x}, y), \tag{2.2}$$

where $\rho\colon\ \mathbb{R}\times\{0,1\}\to\mathbb{R}$ is some function, $\mathbf{b},\mathbf{x}\in\mathbb{R}^{p+1}, y\in\{0,1\}$. In almost all theorems below (except Theorem 2.2) we assume that $\rho(\cdot,y)$ is convex (or strictly convex) function for all y. Differentiability of $\rho(\cdot,y)$ for all y along with LRC is used in Sections 2.2–2.4 to show the form of $\boldsymbol{\beta}^*$ in the fitted model or to show interplay between $\boldsymbol{\beta}^*_{log}$ obtained as minimizer of logistic risk and $\boldsymbol{\beta}^*_{lin}$ obtained as minimizer of quadratic risk. If $\rho(\cdot,y)$ is differentiable for all y, we will denote partial derivative of ρ with respect to first argument by $\frac{\partial\rho}{\partial b}$.

Another important assumption related to uniqueness of $\boldsymbol{\beta}^*$ is linear non-degenerability of $\mathbf{X}$:

(LND) $\mathbb{P}(\mathbf{b}^T\mathbf{X}=0)<1$ for all $\mathbf{b}\in\mathbb{R}^{p+1}\setminus\{\mathbf{0}_{p+1}\}$.

It follows from the Lemma A.44 that if LND is satisfied and q is strictly increasing then we obtain $\mathrm{Cov}(\boldsymbol{\beta}^T\mathbf{X},Y)>0$.

Observe that if $\boldsymbol{\beta}^*$ exists in (1.3), then uniqueness of $\boldsymbol{\beta}^*$ follows easily from strict convexity of $\rho(\cdot,y)$ for all y, LND and condition that for all $\mathbf{b}\in\mathbb{R}^{p+1}$: $\mathbb{E}|\rho(\mathbf{b}^T\mathbf{X},Y)|<\infty$ because risk function R defined in (1.2) is then strictly convex and thus it has unique minimum (see Remark A.1 in Appendix). Note that when assumption LND is not satisfied for a certain $\mathbf{v}\in\mathbb{R}^{p+1}$ and $\boldsymbol{\beta}^*$ is a minimizer of risk function R, then $\boldsymbol{\beta}^*+a\mathbf{v}$ is also minimizer of R for $a\in\mathbb{R}$ as $R(\mathbf{b}+a\mathbf{v})=R(\mathbf{b})$ for every $\mathbf{b}\in\mathbb{R}^{p+1}$. Corollary A.10 in the Appendix gives sufficient conditions for existence of $\boldsymbol{\beta}^*$, however it only works for loss functions of a special form (like logistic and quadratic loss - see Remarks A.12 and A.13). Conditions for existence of $\boldsymbol{\beta}^*$ for e.g. Huber loss and quantile loss remain unknown. Thus existence of $\boldsymbol{\beta}^*$ will be assumed in this chapter in Sections 2.2, 2.3, 2.4 and 2.5.

Condition $\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})\in(0,1)$ $\mathbb{P}_{\mathbf{X}}$ a.e. (assumed in Remark A.12) is crucial for the existence of $\boldsymbol{\beta}^*$ for logistic loss as it is shown in Example 2.40.

Definition 2.1. *We define the following sets of active predictors:*

$$s=\{i\in\{1,\ldots,p\}\colon\ \beta_i\neq 0\}=\operatorname{supp}\tilde{\boldsymbol{\beta}}, \tag{2.3}$$

$$s^*=\{i\in\{1,\ldots,p\}\colon\ \beta^*_i\neq 0\}=\operatorname{supp}\tilde{\boldsymbol{\beta}}^*. \tag{2.4}$$

Note that intercept is not included in s and s^*.

Below we give a few results which also follow from the theorems for the generalized semiparametric setup (see Chapter 3).

2.1. General loss

The following theorem is a generalization of Theorem 3.1 in Kubkowski and Mielniczuk (2017) which was proved there for logistic loss only. Theorem 2.2 states that when inactive predictors $\tilde{\mathbf{X}}_2$ in binary model are such that $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1$ are independent, where $\tilde{\mathbf{X}}_1$ are remaining predictors and $\mathbf{A}$ is a linear transform, then minimizer $\boldsymbol{\beta}^*$ of the risk in the model containing $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ is obtained from the minimizer $(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T})$ of the risk in the model containing only $\tilde{\mathbf{X}}_1$ by appending zeroes to the latter. It is easily seen that the result fails if $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ are dependent (see Example 2.43). Example 2.39 shows possible application of Theorem 2.2.

Theorem 2.2. *Let $\boldsymbol{X}$ be a random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$ and assume that $\mathbb{E}|\rho(\boldsymbol{b}^T\boldsymbol{X}, Y)| < \infty$ for all $\boldsymbol{b} \in \mathbb{R}^{p+1}$. Let $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}_1^T, \tilde{\boldsymbol{X}}_2^T)^T$, where $\tilde{\boldsymbol{X}}_1 = (X_1, \ldots, X_j)^T$, $\tilde{\boldsymbol{X}}_2 = (X_{j+1}, \ldots, X_p)^T$, $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T$, where $\tilde{\boldsymbol{\beta}}_1 \in \mathbb{R}^j, \tilde{\boldsymbol{\beta}}_2 \in \mathbb{R}^{p-j}$. Assume that $\tilde{\boldsymbol{\beta}}_2 = \boldsymbol{0}_{p-j}$ and that $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2 - \boldsymbol{A}\tilde{\boldsymbol{X}}_1$ are independent for a certain $\boldsymbol{A} \in \mathbb{R}^{(p-j)\times j}$. If there exists $(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T})^T$ such that:*

$$(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T})^T = \underset{(b_0, \boldsymbol{b}_1^T)^T \in \mathbb{R}^{j+1}}{\arg\min} \mathbb{E}\rho(b_0 + \boldsymbol{b}_1^T\tilde{\boldsymbol{X}}_1, Y), \tag{2.5}$$

then $\boldsymbol{\beta}^*$ defined in Equation (1.3) exists and:

$$\boldsymbol{\beta}^* = (\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T}, \boldsymbol{0}_{p-j}^T)^T.$$

Moreover, if we assume LND and strict convexity of $\rho(\cdot, y)$ for all y, then $\boldsymbol{\beta}^*$ is unique.

Proof. Let $h(b_0, \mathbf{b}_1^T, \mathbf{b}_2^T) = R(\mathbf{b}) = \mathbb{E}\rho(b_0 + \mathbf{b}_1^T\tilde{\mathbf{X}}_1 + \mathbf{b}_2^T\tilde{\mathbf{X}}_2, Y)$.
Note that h is well specified as $\mathbb{E}|\rho(\mathbf{b}^T\mathbf{X}, Y)| < \infty$ for all $\mathbf{b} \in \mathbb{R}^{p+1}$. Equation (2.5) is equivalent to $h(b_0, \mathbf{b}_1^T, \mathbf{0}_{p-j}^T) \geq h(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T}, \mathbf{0}_{p-j}^T)$ for all $(b_0, \mathbf{b}_1^T)^T \in \mathbb{R}^{j+1}$. Now, by conditioning on $\tilde{\mathbf{X}}$, simple algebraic transformations, conditioning on $\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1$ and independence of $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1$ (last equality) we obtain:

$$\begin{aligned}
h(b_0, \mathbf{b}_1^T, \mathbf{b}_2^T) &= \mathbb{E}\rho(b_0 + \mathbf{b}_1^T\tilde{\mathbf{X}}_1 + \mathbf{b}_2^T\tilde{\mathbf{X}}_2, Y) = \mathbb{E}(\mathbb{E}(\rho(b_0 + \mathbf{b}_1^T\tilde{\mathbf{X}}_1 + \mathbf{b}_2^T\tilde{\mathbf{X}}_2, Y)|\tilde{\mathbf{X}})) \\
&= \mathbb{E}\rho(b_0 + \mathbf{b}_1^T\tilde{\mathbf{X}}_1 + \mathbf{b}_2^T\tilde{\mathbf{X}}_2, 1)q(\boldsymbol{\beta}^T\mathbf{X}) \\
&\quad + \mathbb{E}\rho(b_0 + \mathbf{b}_1^T\tilde{\mathbf{X}}_1 + \mathbf{b}_2^T\tilde{\mathbf{X}}_2, 0)(1 - q(\boldsymbol{\beta}^T\mathbf{X})) \\
&= \mathbb{E}(\rho((b_0 + \mathbf{b}_2^T(\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1)) + (\mathbf{b}_1 + \mathbf{A}^T\mathbf{b}_2)^T\tilde{\mathbf{X}}_1, 1)q(\beta_0 + \tilde{\boldsymbol{\beta}}_1^T\tilde{\mathbf{X}}_1)) \\
&\quad + \mathbb{E}(\rho((b_0 + \mathbf{b}_2^T(\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1)) + (\mathbf{b}_1 + \mathbf{A}^T\mathbf{b}_2)^T\tilde{\mathbf{X}}_1, 0)(1 - q(\beta_0 + \tilde{\boldsymbol{\beta}}_1^T\tilde{\mathbf{X}}_1))) \\
&= \mathbb{E}\Big(\mathbb{E}(\rho((b_0 + \mathbf{b}_2^T(\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1)) + (\mathbf{b}_1 + \mathbf{A}^T\mathbf{b}_2)^T\tilde{\mathbf{X}}_1, 1) \\
&\qquad \times q(\beta_0 + \tilde{\boldsymbol{\beta}}_1^T\tilde{\mathbf{X}}_1)|\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1)\Big) \\
&\quad + \mathbb{E}\Big(\mathbb{E}(\rho((b_0 + \mathbf{b}_2^T(\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1)) + (\mathbf{b}_1 + \mathbf{A}^T\mathbf{b}_2)^T\tilde{\mathbf{X}}_1, 0) \\
&\qquad \times (1 - q(\beta_0 + \tilde{\boldsymbol{\beta}}_1^T\tilde{\mathbf{X}}_1))|\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1)\Big)
\end{aligned}$$

$$= \mathbb{E}h(b_0 + \mathbf{b}_2^T(\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1), \mathbf{b}_1^T + \mathbf{b}_2^T\mathbf{A}, \mathbf{0}_{p-j}^T) \geq h(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T}, \mathbf{0}_{p-j}^T). \tag{2.6}$$

This means that point $\boldsymbol{\beta}^* = (\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T}, \mathbf{0}_{p-j}^T)$ is a global minimum of h. From this the first part of the theorem follows. To prove the second part, we use Lemma A.1 to see that h is strictly convex. This means that $\boldsymbol{\beta}^*$ is unique in view of strict convexity of h. □

Remark 2.3. *If we fit the model without intercept (see Definition 1.8) in the Theorem 2.2, we need additionally to assume $\mathbb{E}\tilde{\boldsymbol{X}} = \boldsymbol{0}_p$ and the proof will be analogous (in the last step before inequality in (2.6) we apply Jensen's inequality):*

$$\begin{aligned} h(0, \boldsymbol{b}_1^T, \boldsymbol{b}_2^T) &= \mathbb{E}h(\boldsymbol{b}_2^T(\tilde{\boldsymbol{X}}_2 - \boldsymbol{A}\tilde{\boldsymbol{X}}_1), \boldsymbol{b}_1^T + \boldsymbol{b}_2^T\boldsymbol{A}, \boldsymbol{0}_{p-j}^T) \\ &\geq h(\boldsymbol{b}_2^T\mathbb{E}(\tilde{\boldsymbol{X}}_2 - \boldsymbol{A}\tilde{\boldsymbol{X}}_1), \boldsymbol{b}_1^T + \boldsymbol{b}_2^T\boldsymbol{A}, \boldsymbol{0}_{p-j}^T) = h(0, \boldsymbol{b}_1^T + \boldsymbol{b}_2^T\boldsymbol{A}, \boldsymbol{0}_{p-j}^T) \geq h(0, \tilde{\boldsymbol{\beta}}_1^{*T}, \boldsymbol{0}_{p-j}^T). \end{aligned}$$

Remark 2.4. *Note that Theorem 2.2 holds in particular when $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$ are independent.*

Corollary 2.5. *If we additionally assume that $\beta_i \neq 0$ for all $i \in \{1, \ldots, j\}$, that is $s = \{1, \ldots, j\}$ then $s^* \subseteq s$ follows from Theorem 2.2. Moreover, we have $\emptyset \neq s^* \subseteq s$, if we assume additionally that $q(x) \in (0, 1)$ for $x \in \mathbb{R}$ (see Remark 2.18).*

The following theorem states that under LRC($\boldsymbol{\beta}$) direction of a $\tilde{\boldsymbol{\beta}}^*$ is the same as direction of $\tilde{\boldsymbol{\beta}}$. This allows us to recover set s from the set s^*. This property was observed for the first time in Brillinger (1982), where $\tilde{\mathbf{X}}$ follows normal distribution and $\boldsymbol{\beta}^*$ is a minimizer of quadratic risk and for loss of the form $-y \ln \pi(\mathbf{x}, \mathbf{b}) - (1 - y) \ln(1 - \pi(\mathbf{x}, \mathbf{b}))$, where $\pi(\mathbf{x}, \mathbf{b}) \in (0, 1)$ in Ruud (1983). The result below is a generalisation of reasoning shown in Section 3 in Ruud (1983), where the assumptions were not given explicitely. For review of similar results we refer to Kubkowski and Mielniczuk (2017). Another proof for $Y = g(\boldsymbol{\beta}^T\mathbf{X}, \varepsilon)$, where ε and $\mathbf{X}$ are independent, can be found in Li and Duan (1989) (see also Remark 2.7). Remark 2.8 shows that semiparametric setup is equivalent to $Y = g(\boldsymbol{\beta}^T\mathbf{X}, \varepsilon)$ considered in Li and Duan (1989) and thus our proof can be considered as certain modification of the proof in Li and Duan (1989). In our version, we do not need to use ε when conditioning expected loss on $\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}$.

Theorem 2.6. *Let $\boldsymbol{X}$ be a random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $\rho(\cdot, y)$ is convex and differentiable function for all y. Assume LRC($\boldsymbol{\beta}$) and that for all $\boldsymbol{b} \in \mathbb{R}^{p+1}$ we have $\mathbb{E}|\rho(\boldsymbol{b}^T\boldsymbol{X}, Y)| < \infty$. If exist $\beta_0^*, \eta \in \mathbb{R}$ such that:*

$$(\beta_0^*, \eta) = \underset{(b_0, c) \in \mathbb{R} \times \mathbb{R}}{\arg\min}\ \mathbb{E}\rho(b_0 + c\tilde{\boldsymbol{\beta}}^T\boldsymbol{X}, Y), \tag{2.7}$$

then $\boldsymbol{\beta}^*$ defined in (1.3) exists and:

$$\boldsymbol{\beta}^* = (\beta_0^*, \eta\tilde{\boldsymbol{\beta}}^T)^T.$$

Moreover, if we assume LND and strict convexity of $\rho(\cdot, y)$ for all y, then $\boldsymbol{\beta}^*$ is unique.

Proof. Let $\mathbf{r} \in \mathbb{R}^p$, $c \in \mathbb{R}$ and $\tilde{\mathbf{b}} = \tilde{\boldsymbol{\beta}} \cdot c + \mathbf{r}$. Then loss l can be written as:

$$l(\mathbf{b}, \mathbf{X}, Y) = \rho(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}, Y) =: h(b_0, c, \mathbf{r}).$$

We define function $J(b_0, c, \mathbf{r}) = \mathbb{E}h(b_0, c, \mathbf{r})$. J is well defined in view of moment assumptions about ρ. We observe that (2.7) is equivalent to: $J(\beta_0^*, \eta, \mathbf{0}_p) \leq J(b_0, c, \mathbf{0}_p)$ for every $b_0 \in \mathbb{R}$ and $c \in \mathbb{R}$. For the first part of the theorem, we need only to show that

$$(\beta_0^*, \eta, \mathbf{0}_p) = \underset{(b_0,c,\mathbf{r})\in\mathbb{R}\times\mathbb{R}\times\mathbb{R}^p}{\arg\min} J(b_0, c, \mathbf{r}). \tag{2.8}$$

Now, by conditioning on $\tilde{\mathbf{X}}$ and then on $\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}$, from LRC($\boldsymbol{\beta}$), Jensen's inequality and (2.7) we obtain:

$$\begin{aligned}
J(b_0, c, \mathbf{r}) &= \mathbb{E}\rho(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}, Y) \\
&= \mathbb{E}(\mathbb{E}(\rho(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}, Y)|\tilde{\mathbf{X}})) \\
&= \mathbb{E}\rho(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}, 1)q(\boldsymbol{\beta}^T\mathbf{X}) \\
&\quad + \mathbb{E}\rho(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}, 0)(1 - q(\boldsymbol{\beta}^T\mathbf{X})) \\
&= \mathbb{E}(\mathbb{E}(\rho(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}, 1)|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}})q(\boldsymbol{\beta}^T\mathbf{X})) \\
&\quad + \mathbb{E}(\mathbb{E}(\rho(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}, 0)|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}})(1 - q(\boldsymbol{\beta}^T\mathbf{X}))) \\
&\geq \mathbb{E}\rho(\mathbb{E}(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}), 1)q(\boldsymbol{\beta}^T\mathbf{X}) \\
&\quad + \mathbb{E}\rho(\mathbb{E}(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}), 0)(1 - q(\boldsymbol{\beta}^T\mathbf{X})) \\
&= \mathbb{E}\rho(\mathbb{E}(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}), Y) \\
&= \mathbb{E}\rho(b_0 + c\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}} + \mathbf{r}^T(\mathbf{h}_0 + \mathbf{h}_1\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}), Y) \\
&= J(b_0 + \mathbf{r}^T\mathbf{h}_0, c + \mathbf{h}_1^T\mathbf{r}, \mathbf{0}_p) \geq J(\beta_0^*, \eta, \mathbf{0}_p).
\end{aligned}$$

This means that point $(\beta_0^*, \eta, \mathbf{0}_p)$ is a global minimum of J. Hence (2.8) is satisfied, $\boldsymbol{\beta}^*$ exists and equals $(\beta_0^*, \eta\tilde{\boldsymbol{\beta}}^T)^T$. Uniqueness of $\boldsymbol{\beta}^*$ is obtained by using similar reasoning as in the proof of the Theorem 2.2. □

Remark 2.7. *Let $Y = g(\boldsymbol{\beta}^T\boldsymbol{X}, \varepsilon)$, where ε and $\boldsymbol{X}$ are independent and g is some function. Original proof of Theorem 2.1 Li and Duan (1989) is the following (we use Jensen's inequality and LRC($\boldsymbol{\beta}$)):*

$$\begin{aligned}
\mathbb{E}\rho(b_0 + \tilde{\boldsymbol{b}}^T\tilde{\boldsymbol{X}}, Y) &= \mathbb{E}(\mathbb{E}(\rho(b_0 + \tilde{\boldsymbol{b}}^T\tilde{\boldsymbol{X}}, g(\boldsymbol{\beta}^T\boldsymbol{X}, \varepsilon)|\boldsymbol{\beta}^T\boldsymbol{X}, \varepsilon)) \\
&\geq \mathbb{E}\rho(b_0 + \mathbb{E}(\tilde{\boldsymbol{b}}^T\tilde{\boldsymbol{X}}|\boldsymbol{\beta}^T\boldsymbol{X}, \varepsilon), Y) = \mathbb{E}\rho(b_0 + \boldsymbol{b}^T\boldsymbol{h}_0 + \boldsymbol{b}^T\boldsymbol{h}_1\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}, Y).
\end{aligned}$$

Remark 2.8. *Let $Y \in \{0, 1\}$, $\mathbb{X} \in \mathbb{R}^{p+1}$. Then the following conditions are equivalent:*

1. *There exists $q : \mathbb{R} \to [0, 1]$ - such that $\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = q(\boldsymbol{\beta}^T\boldsymbol{x})$.*
2. *There exist $g\colon \mathbb{R}^2 \to \{0, 1\}$ and random variable $\varepsilon \in \mathbb{R}$ independent of $\boldsymbol{X}$ such that $Y \stackrel{d}{=} g(\boldsymbol{\beta}^T\boldsymbol{X}, \varepsilon)$.*

Proof. Part 1 follows from part 2, as we have from independence of $\mathbf{X}$ and ε for $\mathbf{x} \in \mathbb{R}^{p+1}$:

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \mathbb{P}(g(\boldsymbol{\beta}^T\mathbf{X}, \varepsilon) = 1|\mathbf{X} = \mathbf{x}) = \mathbb{E}g(\boldsymbol{\beta}^T\mathbf{x}, \varepsilon).$$

This means that $q(\boldsymbol{\beta}^T\mathbf{x}) := \mathbb{E}g(\boldsymbol{\beta}^T\mathbf{x}, \varepsilon)$ satisfies part 1.

Part 2 is implied by part 1, because we take $\varepsilon \sim \mathcal{U}[0,1]$ independent of $\mathbf{X}$, i.e. $\mathbb{P}(\varepsilon \leq t) = t$ for $t \in [0,1]$ and then we have for $\mathbf{x} \in \mathbb{R}^{p+1}$:

$$\mathbb{P}(I(q(\boldsymbol{\beta}^T\mathbf{X}) \geq \varepsilon) = 1|\mathbf{X} = \mathbf{x}) = \mathbb{P}(q(\boldsymbol{\beta}^T\mathbf{x}) \geq \varepsilon) = q(\boldsymbol{\beta}^T\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}).$$

Hence we obtain $Y \stackrel{d}{=} I(q(\boldsymbol{\beta}^T\mathbf{X}) \geq \varepsilon) =: g(\boldsymbol{\beta}^T\mathbf{X}, \varepsilon)$. □

Corollaries 2.9-2.11 allow us to recover set s from s^* for logistic, quadratic and probit loss granted that η introduced in Theorem 2.6 is nonzero.

Corollary 2.9. *If l is logistic loss, $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $q(\boldsymbol{\beta}^T\boldsymbol{X}) \in (0,1)$ $\mathbb{P}_{\boldsymbol{X}}$ a.e., LND and LRC($\boldsymbol{\beta}$) holds then $\boldsymbol{\beta}^* = (\beta_0^*, \eta\boldsymbol{\beta}^T)^T$ for some $\beta_0^*, \eta \in \mathbb{R}$. Moreover $s^* = s$ or $s^* = \emptyset$.*

Proof. From Remark A.12 we obtain that solution of (2.7) exists and moment assumptions of Theorem 2.6 are satisfied. Thus the conclusion follows from Theorem 2.6. □

Analogously, for quadratic and probit losses we obtain in view of Theorem 2.6 (and a note above that theorem) and Remarks A.13-A.14:

Corollary 2.10. *If l is quadratic loss, $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$, $\boldsymbol{\Sigma} > 0$ and LRC($\boldsymbol{\beta}$) holds then $\boldsymbol{\beta}^* = (\beta_0^*, \eta\boldsymbol{\beta}^T)^T$ for some $\beta_0^*, \eta \in \mathbb{R}$. Moreover $s^* = s$ or $s^* = \emptyset$.*

Corollary 2.11. *If l is probit loss, $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$, $\boldsymbol{\Sigma} > 0$, $q(\boldsymbol{\beta}^T\boldsymbol{X}) \in (0,1)$ $\mathbb{P}_{\boldsymbol{X}}$ a.e. and LRC($\boldsymbol{\beta}$) holds then $\boldsymbol{\beta}^* = (\beta_0^*, \eta\boldsymbol{\beta}^T)^T$ for some $\beta_0^*, \eta \in \mathbb{R}$. Moreover $s^* = s$ or $s^* = \emptyset$.*

In the case of the model without intercept, Theorem 2.6 has the following form (with additional assumption $\mathbb{E}\tilde{\mathbf{X}} = \mathbf{0}_p$) and the proof is analogous:

Theorem 2.12. *Let $\boldsymbol{X}$ be a random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $\mathbb{E}\tilde{\boldsymbol{X}} = \boldsymbol{0}_p$, $\rho(\cdot, y)$ is convex and differentiable function for all y. Assume LRC($\boldsymbol{\beta}$) and that for all $\boldsymbol{b} \in \mathbb{R}^{p+1}$ we have $\mathbb{E}|\rho(\boldsymbol{b}^T\boldsymbol{X}, Y)| < \infty$. If there exists $\eta \in \mathbb{R}$ such that:*

$$\eta = \arg\min_{c \in \mathbb{R}} \mathbb{E}\rho(c\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}, Y),$$

then $\boldsymbol{\beta}^*$ defined in (1.8) exists and:

$$\boldsymbol{\beta}^* = \eta\tilde{\boldsymbol{\beta}}.$$

Moreover, if we assume LND and strict convexity of $\rho(\cdot, y)$ for all y, then $\boldsymbol{\beta}^*$ is unique.

Theorem below states conditions under which operations of taking derivative and expectation can be interchanged in

$$D(\mathbb{E}R(\mathbf{b}))_{|\mathbf{b}=\boldsymbol{\beta}^*} = 0.$$

Theorem 2.13. *Assume that $\rho(\cdot, y)$ is differentiable for all y,*

$$\forall \boldsymbol{b} \in \mathbb{R}^{p+1}\colon\ \mathbb{E}|\rho(\boldsymbol{b}^T\boldsymbol{X}, Y)| < \infty,$$

$$\exists h\colon\ \mathbb{R}^{p+1} \times \{0,1\} \to \mathbb{R}\ \forall \boldsymbol{b} \in \mathbb{R}^{p+1}\colon\ \left|\left|\frac{\partial\rho}{\partial b}(\boldsymbol{b}^T\boldsymbol{X}, Y)\boldsymbol{X}\right|\right|_2 \leq h(\boldsymbol{X}, Y)$$

where $\mathbb{E}h(\boldsymbol{X}, Y) < \infty$ *and* $\boldsymbol{\beta}^*$ *exists. Then* $\boldsymbol{\beta}^*$ *is solution to normal equations:*

$$\mathbb{E}\left(\frac{\partial \rho}{\partial b}(\boldsymbol{b}^T\boldsymbol{X}, Y)\boldsymbol{X}\right) = 0. \tag{2.9}$$

Furthermore, if $\rho(\cdot, y)$ *is convex function for all* y *and (2.9) is satisfied for* $\boldsymbol{b} = \boldsymbol{\beta}^*$*, then* $\boldsymbol{\beta}^*$ *is a minimizer of risk function* R*.*

In the case of logistic loss, equality (2.9) reduces further to normal equations for fitting logistic regression:

$$\mathbb{E}(q_L(\boldsymbol{\beta}^{*T}\mathbf{X}) - Y)\mathbf{X} = 0. \tag{2.10}$$

By conditioning on $\mathbf{X}$ in (2.10), we obtain:

$$\mathbb{E}(q_L(\boldsymbol{\beta}^{*T}\mathbf{X}) - q(\boldsymbol{\beta}^T\mathbf{X}))\mathbf{X} = 0. \tag{2.11}$$

In the model with intercept we obtain additionally ($\mathbb{E}Y = \mathbb{E}q_L(\boldsymbol{\beta}^{*T}\mathbf{X})$):

$$\operatorname{Cov}(\tilde{\mathbf{X}}, Y) = \operatorname{Cov}(\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})) = \operatorname{Cov}(\tilde{\mathbf{X}}, q_L(\boldsymbol{\beta}^{*T}\mathbf{X})). \tag{2.12}$$

In the case of probit loss, expression (2.9) reduces to:

$$\mathbb{E}\left(\frac{\phi(\boldsymbol{\beta}^{*T}\mathbf{X})}{\Phi(\boldsymbol{\beta}^{*T}\mathbf{X})(1-\Phi(\boldsymbol{\beta}^{*T}\mathbf{X}))}(\Phi(\boldsymbol{\beta}^{*T}\mathbf{X}) - Y)\mathbf{X}\right) = 0. \tag{2.13}$$

By conditioning on $\mathbf{X}$ in (2.13), we obtain:

$$\mathbb{E}\left(\frac{\phi(\boldsymbol{\beta}^{*T}\mathbf{X})}{\Phi(\boldsymbol{\beta}^{*T}\mathbf{X})(1-\Phi(\boldsymbol{\beta}^{*T}\mathbf{X}))}(\Phi(\boldsymbol{\beta}^{*T}\mathbf{X}) - q(\boldsymbol{\beta}^T\mathbf{X}))\mathbf{X}\right) = 0. \tag{2.14}$$

For quadratic loss we do not need to assume existence of $\boldsymbol{\beta}^*$ - we can replace this assumption by $\mathbb{E}||\tilde{\mathbf{X}}||_2^2 < \infty$ and $\boldsymbol{\Sigma} > 0$. For this loss (2.9) reduces to:

$$\mathbb{E}(\mathbf{X}^T\boldsymbol{\beta}^*_{lin} - Y)\mathbf{X} = 0. \tag{2.15}$$

This means that (after noting that from $\mathbb{E}||\tilde{\mathbf{X}}||_2^2 < \infty$ and $\boldsymbol{\Sigma} > 0$ it follows that $\mathbb{E}\mathbf{X}\mathbf{X}^T$ exists and is invertible as a positive definite matrix):

$$\boldsymbol{\beta}^*_{lin} = (\mathbb{E}\mathbf{X}\mathbf{X}^T)^{-1}\mathbb{E}Y\mathbf{X} = (\mathbb{E}\mathbf{X}\mathbf{X}^T)^{-1}\mathbb{E}q(\boldsymbol{\beta}^T\mathbf{X})\mathbf{X}. \tag{2.16}$$

Normal equations for Huber loss have more complicated form:

$$\frac{1}{\delta}\mathbb{E}(\mathbf{X}\mathbf{X}^T\boldsymbol{\beta}^*_H - Y\mathbf{X})I(|Y - \mathbf{X}^T\boldsymbol{\beta}^*_H| \le \delta)$$
$$- \mathbb{E}\mathbf{X}\operatorname{sgn}(Y - \mathbf{X}^T\boldsymbol{\beta}^*_H)I(|Y - \mathbf{X}^T\boldsymbol{\beta}^*_H| > \delta) = 0. \tag{2.17}$$

Lemma 2.14. *Let* $\boldsymbol{X}$ *be random vector such that* $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$ *with* $\mathbb{E}\tilde{\boldsymbol{X}} = \boldsymbol{\mu}$ *and* $\operatorname{Var}\tilde{\boldsymbol{X}} = \boldsymbol{\Sigma} > 0$*. If* $\boldsymbol{X}$ *satisfies LRC(*$\boldsymbol{\beta}$*) with* $\boldsymbol{h}_0$ *and* $\boldsymbol{h}_1$ *and* $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{0}_p$ *then we have:*

$$\boldsymbol{h}_1 = \frac{\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}}{\tilde{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}}, \tag{2.18}$$

$$\boldsymbol{h}_0 = (\boldsymbol{I}_p - \boldsymbol{h}_1\tilde{\boldsymbol{\beta}}^T)\boldsymbol{\mu} = \left(\boldsymbol{I}_p - \frac{\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^T}{\tilde{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}}\right)\boldsymbol{\mu}. \tag{2.19}$$

Proof. Observe that

$$\operatorname{Cov}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}) = \operatorname{Cov}(\mathbb{E}(\tilde{\mathbf{X}}|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}), \mathbb{E}(\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}})) + \mathbb{E}\operatorname{Cov}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}) \tag{2.20}$$

$$= \operatorname{Cov}(\mathbb{E}(\tilde{\mathbf{X}}|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}), \tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}) + \mathbf{0}_p = \operatorname{Cov}(\mathbf{h}_0 + \mathbf{h}_1\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}, \tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}). \tag{2.21}$$

As $\mathbf{h}_0$ and $\mathbf{h}_1$ are deterministic, we have

$$\operatorname{Cov}(\mathbf{h}_0 + \mathbf{h}_1\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}, \tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}) = \mathbf{h}_1 \operatorname{Var}(\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}) = \mathbf{h}_1\tilde{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}. \tag{2.22}$$

Because $\operatorname{Cov}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}) = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}$, it follows from (2.20) and (2.22) that

$$\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}} = \mathbf{h}_1\tilde{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}$$

and thus $\mathbf{h}_1 = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}})^{-1}$. Formula for $\mathbf{h}_0$ follows after taking expected values of both sides of LRC($\boldsymbol{\beta}$) and using formula for $\mathbf{h}_1$. □

Remark 2.15. *Note that* $\tilde{\boldsymbol{\beta}}^T\boldsymbol{h}_1 = 1$ *and* $\tilde{\boldsymbol{\beta}}^T\boldsymbol{h}_0 = 0$.

2.2. Logistic loss

In this section we assume that function ρ defined in (2.2) is given by the formula:

$$\rho(b, y) = -by + \ln(1 + \exp(b)). \tag{2.23}$$

Here we give another proof of Theorem 2.6 for logistic loss based on normal equations (2.10). The following theorem gives sufficient condition for the proportionality constant η to be nonzero. Method of the proof is based on Brillinger (1982) where the proportionality constant was obtained for linear model. This method allows us to represent $\boldsymbol{\beta}^*$ as a function of $\boldsymbol{\beta}$ even in the situation when some predictors are omitted what is shown in Proposition 2.25.

Theorem 2.16. *Let* $\boldsymbol{X}$ *be random vector such that* $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$ *and* $\boldsymbol{\Sigma} > 0$. *Let* $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{0}_p$. *If* $\boldsymbol{X}$ *satisfies LRC(*$\boldsymbol{\beta}$*) and LRC(*$\boldsymbol{\beta}^*$*) then we have*

$$\tilde{\boldsymbol{\beta}}^* a_{\boldsymbol{\beta}^*} = \tilde{\boldsymbol{\beta}} a_{\boldsymbol{\beta}}, \tag{2.24}$$

where

$$a_{\boldsymbol{\beta}} = (\operatorname{Var}(\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}))^{-1} \operatorname{Cov}(\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}, Y)$$

for $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{0}_p$ *and* $a_{\boldsymbol{\beta}^*}$ *is defined analogously. Moreover if* $\operatorname{Cov}(\boldsymbol{\beta}^T\boldsymbol{X}, Y) \neq 0$, *then* $a_{\boldsymbol{\beta}}, a_{\boldsymbol{\beta}^*} \neq 0$.

Proof. Using covariance decomposition with conditioning vector $\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}$ and Lemma 2.14, we obtain

$$\begin{aligned}
\operatorname{Cov}(\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})) &= \operatorname{Cov}(\mathbb{E}(\tilde{\mathbf{X}}|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}), q(\boldsymbol{\beta}^T\mathbf{X})) + \mathbb{E}\operatorname{Cov}(\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}) = \\
&= \operatorname{Cov}(\mathbf{h}_0 + \mathbf{h}_1\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})) = \\
&= \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}})^{-1} \operatorname{Cov}(\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})) = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}} a_{\boldsymbol{\beta}},
\end{aligned} \tag{2.25}$$

as $\operatorname{Cov}(\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})|\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}) = \mathbf{0}_p$ and the last equality follows from the definition of $a_{\boldsymbol{\beta}}$. Analogously, using linear regressions condition for $\tilde{\boldsymbol{\beta}}^*$ and Lemma 2.14 for $\tilde{\boldsymbol{\beta}}^*$ we obtain the equality:

$$\operatorname{Cov}(\tilde{\mathbf{X}}, q_L(\beta_0^* + \tilde{\mathbf{X}}^T\tilde{\boldsymbol{\beta}}^*)) = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}^* a_{\boldsymbol{\beta}^*}. \tag{2.26}$$

From the normal equations (2.12) we have:

$$\mathrm{Cov}(\tilde{\mathbf{X}}, q_L(\beta_0^* + \tilde{\mathbf{X}}^T\tilde{\boldsymbol{\beta}}^*)) = \mathrm{Cov}(\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})).$$

Thus from (2.25) and (2.26) it follows:

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

From the invertibility of the matrix $\boldsymbol{\Sigma}$ the first part of the theorem follows. To prove the second part, we observe that:

$$0 \neq \mathrm{Cov}(\boldsymbol{\beta}^T\mathbf{X}, Y) = \mathrm{Cov}(\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})) = \tilde{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}} a_{\boldsymbol{\beta}}.$$

This means that $a_{\boldsymbol{\beta}} \neq 0$ (as $\boldsymbol{\Sigma} > 0$ and $\tilde{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}} = ||\boldsymbol{\Sigma}^{\frac{1}{2}}\tilde{\boldsymbol{\beta}}||_2^2$). From the first part of the theorem it follows that $a_{\boldsymbol{\beta}^*} \neq 0$. □

Remark 2.17. *If all of the assumptions imposed in the Theorem 2.16 are satisfied and* $\mathrm{Cov}(\boldsymbol{\beta}^T\boldsymbol{X}, Y) \neq 0$, *then we have* $a_{\boldsymbol{\beta}^*} > 0$. *Namely, it follows from (2.24) that* $\tilde{\boldsymbol{\beta}}^* \neq \boldsymbol{0}_p$ *and in view of Theorem 2.6 we have*

$$\eta = \frac{a_{\boldsymbol{\beta}}}{a_{\boldsymbol{\beta}^*}} \neq 0.$$

From (2.26) and Lemma A.43 we obtain

$$0 \leq \mathrm{Cov}(\tilde{\boldsymbol{\beta}}^{*T}\tilde{\boldsymbol{X}}, q_L(\beta_0^* + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{\beta}}^*)) = \tilde{\boldsymbol{\beta}}^{*T}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}^* a_{\boldsymbol{\beta}^*} = ||\boldsymbol{\Sigma}^{\frac{1}{2}}\tilde{\boldsymbol{\beta}}^*||_2^2 a_{\boldsymbol{\beta}^*}.$$

Thus $a_{\boldsymbol{\beta}^*} > 0$ *as* $a_{\boldsymbol{\beta}^*} \neq 0$ *and we obtain* $s^* = s$.

Remarks 2.18 and 2.19 below show that under some conditions $s^* = \emptyset$ is equivalent to $s = \emptyset$. These results are generalization of Theorem 4 in Mielniczuk and Teisseyre (2016) where assumption about absolute continuity of distribution of $\tilde{\mathbf{X}}$ was imposed. Moreover, in Proposition 1 in Mielniczuk and Teisseyre (2016) it was shown that $s^* = \emptyset$ is equivalent to $\mathbb{E}(\tilde{\mathbf{X}}|Y = 1) = \mathbb{E}(\tilde{\mathbf{X}}|Y = 0)$, if $\mathbb{E}||\tilde{\mathbf{X}}||_2 < \infty$.

Remark 2.18. *If* $\boldsymbol{X}$ *is a random vector satisfying LND,* $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $q(z) \in (0, 1)$ *for* $z \in \mathbb{R}$ *and* $\tilde{\boldsymbol{\beta}} = \boldsymbol{0}_p$, *then* $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{0}_p$.
To prove this, we observe that normal equations (2.10) imply that:

$$\mathrm{Cov}(\tilde{\boldsymbol{X}}, q_L(\boldsymbol{\beta}^{*T}\boldsymbol{X})) = \mathrm{Cov}(\tilde{\boldsymbol{X}}, q(\beta_0)) = \boldsymbol{0}_p.$$

This means that $\mathrm{Cov}(\boldsymbol{\beta}^{*T}\boldsymbol{X}, q_L(\boldsymbol{\beta}^{*T}\boldsymbol{X})) = 0$. *In view of Lemma A.44 we have* $\mathbb{P}(\boldsymbol{\beta}^{*T}\boldsymbol{X} = c) = 1$ *for some* $c \in \mathbb{R}$. *This equality and LND together imply that* $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{0}_p$.

Remark 2.19. *If* $\boldsymbol{X}$ *is a random vector satisfying LND,* $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $q(z) \in (0, 1)$ *for* $z \in \mathbb{R}$, q *is strictly monotone and* $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{0}_p$, *then* $\tilde{\boldsymbol{\beta}} = \boldsymbol{0}_p$.
Proof of this fact is analogous to proof of Remark 2.18.

Remark 2.20. *Theorem 2.16 holds also for the model without intercept. We have in this case:*

$$a_{\boldsymbol{\beta}} = (\mathrm{Var}(\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}))^{-1}\mathbb{E}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}Y$$

and

$$a_{\boldsymbol{\beta}^*} = (\mathrm{Var}(\tilde{\boldsymbol{\beta}}^{*T}\tilde{\boldsymbol{X}}))^{-1}\mathbb{E}\tilde{\boldsymbol{\beta}}^{*T}\tilde{\boldsymbol{X}}Y.$$

Lemma 2.21. *Assume that* $\mathbb{E}\tilde{\boldsymbol{X}} = \boldsymbol{0}_p$, $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $\boldsymbol{X}$ *satisfies LND,* $\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) \in (0, 1)$ $\mathbb{P}_{\boldsymbol{X}}$ *a.e. If* $\boldsymbol{\beta}^*, \boldsymbol{\beta}^*_{-0}$ *denote respectively argmins of risk function* R *in models with intercept and without intercept,* s^*, s^*_{-0} *are sets of active predictors corresponding to them, then* $s^* = \emptyset$ *if and only if* $s^*_{-0} = \emptyset$.

Proof. If $s^* = \emptyset$, then in view of normal equations for the model with intercept we have:

$$\mathbb{E}Y\tilde{\mathbf{X}} = \mathbb{E}q_L(\beta_0^*)\tilde{\mathbf{X}} = \mathbf{0}_p.$$

Now, in view of normal equations for the model without intercept and from above equation we obtain:

$$\mathbb{E}q_L(\boldsymbol{\beta}^{*T}_{-0}\tilde{\mathbf{X}})\tilde{\mathbf{X}} = \mathbb{E}Y\tilde{\mathbf{X}} = \mathbf{0}_p.$$

This implies:

$$\mathbb{E}q_L(\boldsymbol{\beta}^{*T}_{-0}\tilde{\mathbf{X}})\boldsymbol{\beta}^{*T}_{-0}\tilde{\mathbf{X}} = 0.$$

Using Lemma A.44 yields $\mathbb{P}(\boldsymbol{\beta}^{*T}_{-0}\tilde{\mathbf{X}} = c) = 1$ for some $c \in \mathbb{R}$. From the equality above it follows that:

$$c = \mathbb{E}\boldsymbol{\beta}^{*T}_{-0}\tilde{\mathbf{X}} = 0.$$

From LND condition we obtain $\boldsymbol{\beta}^*_{-0} = \mathbf{0}_p$. Hence $s^*_{-0} = \emptyset$.

Proof of implication $s^*_{-0} = \emptyset \Rightarrow s^* = \emptyset$ is analogous. □

From the above lemma, Corollary 2.9 and Theorems 2.6 and 2.12 follows the following remark which states that active sets of predictors for logistic models with intercept and without intercept are always the same under appropriate assumptions.

Remark 2.22. *If assumptions of Lemma 2.21 are satisfied and* $\boldsymbol{X}$ *satisfies LRC(*$\boldsymbol{\beta}$*), then* $s^* = s^*_{-0} = s$ *or* $s^* = s^*_{-0} = \emptyset$.

If $\mathbf{X}$ follows normal distribution, formulas for $a_{\boldsymbol{\beta}}$ and $a_{\boldsymbol{\beta}^*}$ can be simplified.

Lemma 2.23. *If* q *is differentiable,* $\tilde{\boldsymbol{X}}$ *follows normal distribution with* $\boldsymbol{\Sigma} > 0$, $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{0}_p$, $\mathbb{E}|q'(\boldsymbol{\beta}^T\boldsymbol{X})| < \infty$, *then* $a_{\boldsymbol{\beta}} = \mathbb{E}q'(\boldsymbol{\beta}^T\boldsymbol{X})$ *and* $a_{\boldsymbol{\beta}^*} = \mathbb{E}q'_L(\boldsymbol{\beta}^{*T}\boldsymbol{X}) \in (0, 1/4)$.

Proof. From Lemma A.45 we obtain

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\beta}^T\mathbf{X})a_{\boldsymbol{\beta}} &= \mathrm{Cov}(\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}, Y) = \mathrm{Cov}(\tilde{\boldsymbol{\beta}}^T\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})) = \mathrm{Cov}(\boldsymbol{\beta}^T\mathbf{X}, q(\boldsymbol{\beta}^T\mathbf{X}))\\ &= \mathrm{Var}(\boldsymbol{\beta}^T\mathbf{X})\mathbb{E}q'(\boldsymbol{\beta}^T\mathbf{X}).\end{aligned}$$

Proof for $a_{\boldsymbol{\beta}^*}$ is analogous (we use there additionally normal equations for logistic loss (2.12)). Moreover, $a_{\boldsymbol{\beta}^*} \in (0, 1/4)$ follows from

$$q'_L(x) = \frac{e^x}{(1+e^x)^2} = q_L(x)(1 - q_L(x)) \in (0, 1/4) \quad \text{for } x \neq 0.$$ □

Remark 2.24. *If all of assumptions of Lemma 2.23 are satisfied then*

$$\eta = \frac{\mathbb{E}q'(\boldsymbol{\beta}^T\boldsymbol{X})}{\mathbb{E}q'_L(\boldsymbol{\beta}^{*T}\boldsymbol{X})} \tag{2.27}$$

and

$$|\eta| > 4|\mathbb{E}q'(\boldsymbol{\beta}^T\boldsymbol{X})|. \tag{2.28}$$

Proposition 2.25. *Let* $\boldsymbol{X} = (X_0, \tilde{\boldsymbol{X}}_1^T, \tilde{\boldsymbol{X}}_2^T)^T$, $\boldsymbol{\beta} = (\beta_0, \tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T$ *and* $\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{X}}_1 \in \mathbb{R}^m$, $\tilde{\boldsymbol{\beta}}_2, \tilde{\boldsymbol{X}}_2 \in \mathbb{R}^{p-m}$, $\operatorname{Cov}(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{X}}_j) = \boldsymbol{\Sigma}_{ij}$ *for* $i, j = 1, 2$. *Suppose that logistic model* $Y \sim b_0 + \tilde{\boldsymbol{X}}_1^T\tilde{\boldsymbol{b}}_1$ *with omitted* $\tilde{\boldsymbol{X}}_2$ *is fitted and* $(\beta_0^*, \boldsymbol{\beta}_1^{*T})^T$ *is the corresponding projection. If* $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$, $\boldsymbol{\Sigma}_{11} > 0$, *LRC(*$\boldsymbol{\beta}$*) and LRC(*$(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T})^T$*) hold and* $\operatorname{Cov}(\boldsymbol{\beta}^T\boldsymbol{X}, Y) \neq 0$, *then we have*

$$\tilde{\boldsymbol{\beta}}_1^* = \eta(\tilde{\boldsymbol{\beta}}_1 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\tilde{\boldsymbol{\beta}}_2), \tag{2.29}$$

where $\eta = \frac{a_{\boldsymbol{\beta}}}{a_{\boldsymbol{\beta}_1^*}} \neq 0$ *and*

$$a_{\boldsymbol{\beta}_1^*} = \frac{\operatorname{Cov}(Y, \tilde{\boldsymbol{X}}_1^T\tilde{\boldsymbol{\beta}}_1^*)}{\operatorname{Var}(\tilde{\boldsymbol{X}}_1^T\tilde{\boldsymbol{\beta}}_1^*)} = \frac{\operatorname{Cov}(q_L(\beta_0^* + \tilde{\boldsymbol{X}}_1^T\tilde{\boldsymbol{\beta}}_1^*), \tilde{\boldsymbol{X}}_1^T\tilde{\boldsymbol{\beta}}_1^*)}{\operatorname{Var}(\tilde{\boldsymbol{X}}_1^T\tilde{\boldsymbol{\beta}}_1^*)}. \tag{2.30}$$

Proof. Analogously as in Theorem 2.16, we obtain the equations:

$$\operatorname{Cov}(\tilde{\mathbf{X}}_1, q(\beta_0 + \tilde{\boldsymbol{\beta}}_1^T\tilde{\mathbf{X}}_1 + \tilde{\boldsymbol{\beta}}_2^T\tilde{\mathbf{X}}_2)) = a_{\boldsymbol{\beta}}\operatorname{Cov}(\tilde{\mathbf{X}}_1, \tilde{\boldsymbol{\beta}}_1^T\tilde{\mathbf{X}}_1 + \tilde{\boldsymbol{\beta}}_2^T\tilde{\mathbf{X}}_2),$$
$$\operatorname{Cov}(\tilde{\mathbf{X}}_1, q_L(\beta_0^* + \tilde{\boldsymbol{\beta}}_1^{*T}\tilde{\mathbf{X}}_1)) = a_{\boldsymbol{\beta}_1^*}\operatorname{Cov}(\tilde{\mathbf{X}}_1, \tilde{\boldsymbol{\beta}}_1^{*T}\tilde{\mathbf{X}}_1).$$

Hence from normal equations we obtain

$$a_{\boldsymbol{\beta}_1^*}\operatorname{Cov}(\tilde{\mathbf{X}}_1, \tilde{\boldsymbol{\beta}}_1^*\tilde{\mathbf{X}}_1) = a_{\boldsymbol{\beta}}\operatorname{Cov}(\tilde{\mathbf{X}}_1, \tilde{\boldsymbol{\beta}}_1^T\tilde{\mathbf{X}}_1 + \tilde{\boldsymbol{\beta}}_2^T\tilde{\mathbf{X}}_2),$$

what can be simplified to

$$a_{\boldsymbol{\beta}_1^*}\boldsymbol{\Sigma}_{11}\tilde{\boldsymbol{\beta}}_1^* = a_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}_{11}\tilde{\boldsymbol{\beta}}_1 + \boldsymbol{\Sigma}_{12}\tilde{\boldsymbol{\beta}}_2). \tag{2.31}$$

As $\boldsymbol{\Sigma}_{11}$ is invertible, we conclude that $a_{\boldsymbol{\beta}_1^*}$ is non-zero similarly as in Theorem 2.16. Multiplying both sides of Equation (2.31) by $(a_{\boldsymbol{\beta}_1^*}\boldsymbol{\Sigma}_{11})^{-1}$, we obtain the conclusion. □

Remark 2.26. *If in Proposition 2.25 we assume additionally that* $\boldsymbol{\Sigma}_{12} = \boldsymbol{O}_{m\times(p-m)}$, *then* $\tilde{\boldsymbol{\beta}}_1^* = \eta\tilde{\boldsymbol{\beta}}_1$. *For independent* $\tilde{\boldsymbol{X}}_1$ *and* $\tilde{\boldsymbol{X}}_2$ *we thus obtain a complementary conclusion to that of Theorem 2.2.*

2.3. Quadratic loss

In this section we assume that function ρ defined in (2.2) is given by the formula:

$$\rho(b, y) = \frac{1}{2}(y - b)^2. \tag{2.32}$$

Here we give another proof of Theorem 2.6 for quadratic loss based on normal equations (2.15). The following theorem gives sufficient condition for the proportionality constant η being nonzero and provides explicit formula for $\boldsymbol{\beta}^*$. The proof is similar to that of Theorem 2.16. This method allows us to represent $\boldsymbol{\beta}^*$ as a function of $\boldsymbol{\beta}$ even in the situation when some predictors are omitted what is shown in Proposition 2.32.

Theorem 2.27. *Let $\boldsymbol{X}$ be random vector with $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$ and $\boldsymbol{\Sigma} > 0$. If LRC($\boldsymbol{\beta}$) is satisfied then*

$$\tilde{\boldsymbol{\beta}}^* = \tilde{\boldsymbol{\beta}} a_{\boldsymbol{\beta}}, \tag{2.33}$$

$$\beta_0^* = \mathbb{E}q(\boldsymbol{\beta}^T \boldsymbol{X}) - \boldsymbol{\mu}^T \tilde{\boldsymbol{\beta}} a_{\boldsymbol{\beta}}, \tag{2.34}$$

where $a_{\boldsymbol{\beta}}$ is defined similarly as in the Theorem 2.16. Moreover if $\mathrm{Cov}(\boldsymbol{\beta}^T \boldsymbol{X}, Y) \neq 0$, then $\tilde{\boldsymbol{\beta}}^*, a_{\boldsymbol{\beta}} \neq 0$.

Proof. From (2.15) we obtain:

$$\mathbb{E}\mathbf{X}\mathbf{X}^T\boldsymbol{\beta}^* = \mathbb{E}\mathbf{X}Y.$$

This means that we obtain system of linear equations:

$$\begin{cases} \beta_0^* + \mathbb{E}\tilde{\mathbf{X}}^T\tilde{\boldsymbol{\beta}}^* = \mathbb{E}q(\boldsymbol{\beta}^T\mathbf{X}), \\ \mathbb{E}\tilde{\mathbf{X}}\beta_0^* + \mathbb{E}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\tilde{\boldsymbol{\beta}}^* = \mathbb{E}\tilde{\mathbf{X}}q(\boldsymbol{\beta}^T\mathbf{X}). \end{cases} \tag{2.35}$$

Hence, using (2.25), we have:

$$\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}^* = \mathrm{Var}\,\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}^* = \mathrm{Cov}(\tilde{\mathbf{X}}, q(\boldsymbol{\beta}^T\mathbf{X})) = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}a_{\boldsymbol{\beta}}.$$

Matrix $\boldsymbol{\Sigma}$ is invertible as it is positive definite and we obtain:

$$\tilde{\boldsymbol{\beta}}^* = \tilde{\boldsymbol{\beta}}a_{\boldsymbol{\beta}},$$

Thus, after substituting $\tilde{\boldsymbol{\beta}}^*$ into (2.35), we have:

$$\beta_0^* = \mathbb{E}q(\boldsymbol{\beta}^T\mathbf{X}) - \mathbb{E}\tilde{\mathbf{X}}^T\tilde{\boldsymbol{\beta}}^* = \mathbb{E}q(\boldsymbol{\beta}^T\mathbf{X}) - \boldsymbol{\mu}^T\tilde{\boldsymbol{\beta}}a_{\boldsymbol{\beta}}.$$

Second part of the proof is identical as in proof of the Theorem 2.16. □

Remark 2.28. *Theorem 2.27 holds also for the model without intercept. We have in this case (see Remark 2.20):*

$$a_{\boldsymbol{\beta}} = (\mathrm{Var}(\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}))^{-1}\mathbb{E}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}Y.$$

The following lemma is a version of Lemma 2.21 for quadratic loss.

Lemma 2.29. *Assume that $\mathbb{E}\tilde{\boldsymbol{X}} = \boldsymbol{0}_p$, $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$ and $\boldsymbol{\Sigma} = \mathrm{Var}(\tilde{\boldsymbol{X}}) > 0$. If $\boldsymbol{\beta}^*, \boldsymbol{\beta}_{-0}^*$ denote respectively argmins of risk function R in models with intercept and without intercept, s^*, s_{-0}^* are corresponding to them sets of active predictors, then $s^* = \emptyset$ if and only if $s_{-0}^* = \emptyset$.*

Proof. If $s^* = \emptyset$, then in view of normal equations for the model with intercept we have:

$$\mathbb{E}Y\tilde{\mathbf{X}} = \mathbb{E}\beta_0^*\tilde{\mathbf{X}} = \mathbf{0}_p.$$

Now, in view of normal equations for the model without intercept and from above equation we obtain:

$$\mathbb{E}\boldsymbol{\beta}_{-0}^{*T}\tilde{\mathbf{X}}\tilde{\mathbf{X}} = \mathbb{E}Y\tilde{\mathbf{X}} = \mathbf{0}_p.$$

This means that:

$$\mathbb{E}(\boldsymbol{\beta}_{-0}^{*T}\tilde{\mathbf{X}}) \cdot (\boldsymbol{\beta}_{-0}^{*T}\tilde{\mathbf{X}}) = 0.$$

Using Lemma A.44 yields $\mathbb{P}(\boldsymbol{\beta}_{-0}^{*T}\tilde{\mathbf{X}} = c) = 1$ for some $c \in \mathbb{R}$. We observe that:

$$c = \mathbb{E}\boldsymbol{\beta}_{-0}^{*T}\tilde{\mathbf{X}} = 0.$$

From LND condition we obtain $\boldsymbol{\beta}_{-0}^{*} = \mathbf{0}_p$. Hence $s_{-0}^{*} = \emptyset$.

Proof of implication $s_{-0}^{*} = \emptyset \Rightarrow s^{*} = \emptyset$ is analogous. □

From the above lemma, Corollary 2.10 and Theorems 2.6 and 2.12 follows the following remark which states that active sets of predictors for logistic models with intercept and without intercept are always the same under appropriate assumptions.

Remark 2.30. *If assumptions of Lemma 2.29 are satisfied and $\boldsymbol{X}$ satisfies LRC($\boldsymbol{\beta}$), then $s^{*} = s_{-0}^{*} = s$ or $s^{*} = s_{-0}^{*} = \emptyset$.*

Lemma 2.31. *If q is differentiable, $\tilde{\boldsymbol{X}}$ follows normal distribution, $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$, $\boldsymbol{\Sigma} > 0$ and $\mathbb{E}|q'(\boldsymbol{\beta}^T\boldsymbol{X})| < \infty$, then $a_{\boldsymbol{\beta}} = \mathbb{E}q'(\boldsymbol{\beta}^T\boldsymbol{X})$.*

Proof. Proof is analogous to that of Lemma 2.23. □

Proposition 2.32. *Let $\boldsymbol{X} = (X_0, \tilde{\boldsymbol{X}}_1^T, \tilde{\boldsymbol{X}}_2^T)^T$, $\boldsymbol{\beta} = (\beta_0, \tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T$ and $\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{X}}_1 \in \mathbb{R}^m$, $\tilde{\boldsymbol{\beta}}_2, \tilde{\boldsymbol{X}}_2 \in \mathbb{R}^{p-m}$, $\mathrm{Cov}(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{X}}_j) = \boldsymbol{\Sigma}_{ij}$ for $i, j = 1, 2$. Suppose that logistic model $Y \sim b_0 + \tilde{\boldsymbol{X}}_1^T\tilde{\boldsymbol{b}}_1$ with omitted $\tilde{\boldsymbol{X}}_2$ is fitted and $(\beta_0^{*}, \boldsymbol{\beta}_1^{*T})^T$ is the corresponding projection. Under assumptions of Theorem 2.27 for $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}_1^{*}$ and provided that $\mathrm{Cov}(Y, \boldsymbol{\beta}^T\boldsymbol{X}) \neq 0$ we have*

$$\tilde{\boldsymbol{\beta}}_1^{*} = a_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}_1 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\tilde{\boldsymbol{\beta}}_2), \tag{2.36}$$

where $a_{\boldsymbol{\beta}} \neq 0$.

Proof. Proof is similar to the proof of Proposition 2.25. □

2.4. Quadratic loss vs logistic loss

In this section we compare vectors

$$\boldsymbol{\beta}_{log}^{*} = (\beta_{0,log}^{*}, \tilde{\boldsymbol{\beta}}_{1,log}^{*T}, \tilde{\boldsymbol{\beta}}_{2,log}^{*T})^T = \underset{\mathbf{b}=(b_0,\mathbf{b}_1^T,\mathbf{b}_2^T)^T:\mathbf{b}_2=0}{\arg\min} \mathbb{E}l_{log}(\mathbf{b}, \mathbf{X}, Y),$$

$$\boldsymbol{\beta}_{lin}^{*} = (\beta_{0,lin}^{*}, \tilde{\boldsymbol{\beta}}_{1,lin}^{*T}, \tilde{\boldsymbol{\beta}}_{2,lin}^{*T})^T = \underset{\mathbf{b}=(b_0,\mathbf{b}_1^T,\mathbf{b}_2^T)^T:\mathbf{b}_2=0}{\arg\min} \mathbb{E}l_{lin}(\mathbf{b}, \mathbf{X}, Y)$$

and sets of active predictors corresponding to them:

$$s_{log}^{*} = \operatorname{supp}\tilde{\boldsymbol{\beta}}_{log}^{*},$$

$$s_{lin}^{*} = \operatorname{supp}\tilde{\boldsymbol{\beta}}_{lin}^{*}.$$

It turns out that $\tilde{\boldsymbol{\beta}}_{1,log}^{*}$ and $\tilde{\boldsymbol{\beta}}_{1,lin}^{*}$ will have the same direction under linear regressions conditions. Proposition 2.33 shows that result holds even when in the fitted logistic model we omit predictors $\tilde{\mathbf{X}}_2$ from the vector $\mathbf{X} = (X_0, \tilde{\mathbf{X}}_1^T, \tilde{\mathbf{X}}_2^T)^T$.

Proposition 2.33. *Let $\boldsymbol{X}$ be a random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$ and LRC is satisfied for $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{\beta}}^*_{1,log}$. We assume that $\mathbb{E}||\tilde{\boldsymbol{X}}_1||_2^2 < \infty$ and $\operatorname{Var}\tilde{\boldsymbol{X}}_1 = \boldsymbol{\Sigma}_{11} > 0$. If $\boldsymbol{\beta}^*_{log}$ exists then:*

$$\tilde{\boldsymbol{\beta}}^*_{1,lin} = a_{\beta^*_{1,log}}\tilde{\boldsymbol{\beta}}^*_{1,log}. \tag{2.37}$$

Proof. Observe that:

$$\operatorname{Cov}(\tilde{\mathbf{X}}_1, q_L(\beta^*_{0,log} + \tilde{\boldsymbol{\beta}}^{*T}_{1,log}\tilde{\mathbf{X}}_1)) = \operatorname{Cov}(\tilde{\mathbf{X}}_1, Y) = \operatorname{Cov}(\tilde{\mathbf{X}}_1, \beta^*_{0,lin} + \tilde{\boldsymbol{\beta}}^{*T}_{1,lin}\tilde{\mathbf{X}}_1) = \boldsymbol{\Sigma}_{11}\tilde{\boldsymbol{\beta}}^*_{1,lin}.$$

Moreover, reasoning as in proof of Theorem 2.16 we have:

$$\operatorname{Cov}(\tilde{\mathbf{X}}_1, q_L(\beta^*_{0,log} + \tilde{\boldsymbol{\beta}}^{*T}_{1,log}\tilde{\mathbf{X}}_1)) = a_{\beta^*_{1,log}} \cdot \boldsymbol{\Sigma}_{11}\tilde{\boldsymbol{\beta}}^*_{1,log}.$$

Thus from two last equalities we obtain matrix equation

$$\boldsymbol{\Sigma}_{11}\tilde{\boldsymbol{\beta}}^*_{1,lin} = a_{\beta^*_{1,log}}\boldsymbol{\Sigma}_{11}\tilde{\boldsymbol{\beta}}^*_{1,log},$$

which is equivalent to (2.37). □

Remark 2.34. *In particular the result hold for $\tilde{\boldsymbol{X}}_1 = \tilde{\boldsymbol{X}}$ when all regressors are fitted and for $\tilde{\boldsymbol{X}}_1 = X_j$ when the univariate regressor X_j is fitted. In the latter case linear regression condition for X_j and $\tilde{\boldsymbol{\beta}}^*_{1,log}$ is always satisfied. Note also that when model is correctly specified as logistic model and $\tilde{\boldsymbol{X}}_1 = \tilde{\boldsymbol{X}}$, it follows that $\tilde{\boldsymbol{\beta}}^*_{lin}$ is proportional to $\tilde{\boldsymbol{\beta}}$. 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 2.35. *If $\tilde{\boldsymbol{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then from Lemma 2.23 we obtain:*

$$a_{\beta^*_{1,log}} = \mathbb{E}q'_L(\boldsymbol{\beta}^{*T}_{1,log}\tilde{\boldsymbol{X}}_1) \in (0, 1/4).$$

In view of Proposition 2.33 it follows that corresponding coefficients of vectors $\tilde{\boldsymbol{\beta}}^*_{1,log}$ and $\tilde{\boldsymbol{\beta}}^*_{1,lin}$ have identical signs and $|\tilde{\boldsymbol{\beta}}^*_{1,lin,i}| < \frac{1}{4}|\tilde{\boldsymbol{\beta}}^*_{1,log,i}|$. As $a_{\beta^*_{1,log}} > 0$, we obtain equality $s^*_{1,lin} = s^*_{1,log}$, where $s^*_{1,log} = \operatorname{supp}\tilde{\boldsymbol{\beta}}^*_{1,log}$ and $s^*_{1,lin} = \operatorname{supp}\tilde{\boldsymbol{\beta}}^*_{1,lin}$. In particular, when $\tilde{\boldsymbol{X}}_1 = \tilde{\boldsymbol{X}}$ we have $s^*_{lin} = s^*_{log}$.

2.5. Sets of active predictors when LRC is not imposed

In this section we consider various sets of predictors in the regression problem for $\mathbf{Z} = (1, \tilde{\mathbf{Z}}^T)^T \in \mathbb{R}^{p+1}$ and $V \in \{0, 1\}$ satisfying relation

$$\mathbb{P}(V = 1|\mathbf{Z}) = q(\boldsymbol{\beta}^T(\mathbf{Z})\mathbf{Z})$$

for fixed vector

$$\boldsymbol{\beta}(\mathbf{Z}) = (\beta_0(\mathbf{Z}), \tilde{\boldsymbol{\beta}}(\mathbf{Z})^T)^T \in \mathbb{R}^{p+1}.$$

We write $\boldsymbol{\beta} = \boldsymbol{\beta}(\mathbf{Z})$ to underline the fact that regression parameter $\boldsymbol{\beta}$ is considered in a model with predictors $\mathbf{Z}$. We will denote by

$$\boldsymbol{\beta}^*(\mathbf{Z}) = (\beta^*_0(\mathbf{Z}), \tilde{\boldsymbol{\beta}}^*(\mathbf{Z})^T)^T = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^{p+1}} \mathbb{E}\rho(\mathbf{b}^T\mathbf{Z}, V)$$

coefficients of a fitted model with predictors $\mathbf{Z}$ and by

$$s^*(\mathbf{Z}) = \operatorname{supp} \tilde{\boldsymbol{\beta}}^*(\mathbf{Z}),$$
$$s(\mathbf{Z}) = \operatorname{supp} \tilde{\boldsymbol{\beta}}(\mathbf{Z}).$$

In this section we assume that function ρ defined in (2.2) is given by the formula:

$$\rho(b, y) = -by + \ln(1 + \exp(b)). \tag{2.38}$$

Lemma 2.36. *Let $q\colon\ \mathbb{R} \to (0,1)$ be uniformly continuous function, $\boldsymbol{Z}_m = (1, \tilde{\boldsymbol{Z}}_m^T)^T$,*

$$\tilde{\boldsymbol{Z}}_m = (Z_{m1}, \ldots, Z_{mp})^T \sim \sum_{l=1}^{p} p_l \mathcal{N}_p(\boldsymbol{x}_l, \frac{\sigma^2}{m} I_p),$$

where $l = 1, \ldots, p$, $p_l > 0$, $\sum\limits_{l=1}^{p} p_l = 1$, $x_l \in \mathbb{R}^p$, $\sigma > 0$. Let $\mathbb{P}(\boldsymbol{Z} = \boldsymbol{x}_l) = p_l$, $l = 1, \ldots, p$ and assume that $\mathbb{P}(\boldsymbol{b}^T \boldsymbol{Z} = 0) < 1$ for all $\boldsymbol{b} \in \mathbb{R}^{p+1}$. If $s^*(\boldsymbol{Z}) = \{1, \ldots, p\}$, then $s^*(\boldsymbol{Z}_m) = s^*(\boldsymbol{Z})$ for sufficiently large m and uniformly continuous function q.

Proof. Firstly, it should be noted that $\boldsymbol{\beta}^*(\mathbf{Z}_m)$ and $\boldsymbol{\beta}^*(\mathbf{Z})$ exist and are unique (see Remark A.12).
By Theorem A.49 we have $\boldsymbol{\beta}^*(\mathbf{Z}_m) \to \boldsymbol{\beta}^*(\mathbf{Z})$ and moreover we know that $s^*(\mathbf{Z}) = \{1, \ldots, p\}$. Thus for all $i = 1, \ldots, p$ we have $\beta_i^*(\mathbf{Z}) \neq 0$, and hence for sufficiently large m we have $\beta_i^*(\mathbf{Z}_m) \neq 0$. □

Lemma 2.37. *If $p \in \mathbb{N}$, $p \geq 2$, $k \in \{1, \ldots, p-1\}$ and $q\colon\ \mathbb{R} \to (0,1)$ is continuous response function such that model is misspecified with respect to logistic loss (i.e. for all $a, b \in \mathbb{R}$ there exists $x \in \mathbb{R}:\ q(x) \neq q_L(ax + b)$), then there exists random vector $\boldsymbol{Z}$ such that $s^*(\boldsymbol{Z}) = \{1, \ldots, p\}$, $s(\boldsymbol{Z}) = \{1, \ldots, k\}$ and $\mathbb{P}(\boldsymbol{b}^T \boldsymbol{Z} = 0) < 1$ for all $\boldsymbol{b} \in \mathbb{R}^{p+1}$.*

Proof. Let us define $f(x) = q_L^{-1}(q(x))$ which by assumption about misspecification of logistic model is a nonlinear function. Our goal is to define linearly non-degenerate random vector $\mathbf{Z} = (1, Z_1, \ldots, Z_p)^T$ such that:

$$Z_1 + \ldots + Z_p = f(Z_1 + \ldots + Z_k), \tag{2.39}$$

that is

$$q_L(Z_1 + \ldots + Z_p) = q(Z_1 + \ldots + Z_k).$$

Then it is obvious that with $\beta_i(\mathbf{Z}) = I(i \leq k)$, this implies $\beta_i^*(\mathbf{Z}) = 1$ for $i = 1, \ldots, p$ and thus $s(\mathbf{Z}) = \{1, \ldots, k\}$ and $s^*(\mathbf{Z}) = \{1, \ldots, p\}$. To this end let $\Omega = \{1, \ldots, p+1\}$ and $\mathbb{P}(\{j\}) = 1/(p+1)$ for all $j = 1, \ldots, p+1$.

For $u_1, u_2 \in \mathbb{R}$ specified later define matrix:

$$\mathbf{A} = [a_{i_1,i_2}]_{1\le i_1,i_2\le p+1} = \begin{bmatrix} 1 & 1 & \cdots & 0 & 0 & \cdots & 0 & f(1)-1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 & 0 & \cdots & 0 & f(1)-1 \\ 1 & 0 & \cdots & 0 & 1 & \cdots & 0 & f(0)-1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & 0 & \cdots & 1 & f(0)-1 \\ 1 & u_1 & 0 & \cdots & \cdots & \cdots & 0 & f(u_1)-u_1 \\ 1 & u_2 & 0 & \cdots & \cdots & \cdots & 0 & f(u_2)-u_2 \end{bmatrix}$$

and $Z_j(i) = a_{i,j+1}$ for $i = 1, \ldots, p+1$ and $j = 0, 1, \ldots, p$. Then $\mathbf{Z} = (Z_0, Z_1, \ldots, Z_p)^T$ satisfies (2.39). Now we will choose u_1, u_2 such that $\mathbf{Z}$ is linearly non-degenerate. From the definition of $\mathbf{Z}$ this condition is equivalent to existence of $\mathbf{b} \in \mathbb{R}^{p+1} \setminus \{\mathbf{0}_{p+1}\}$ that system of linear equations $\mathbf{Ab} = \mathbf{0}_{p+1}$ has nonzero solution.

We observe that $|\det \mathbf{A}| = |(1-u_2)(f(u_1) - u_1 f(1)) - (1-u_1)(f(u_2) - u_2 f(1))|$. We will prove that we can choose u_1, u_2 that this determinant is nonzero and the theorem will follow. From nonlinearity of f there exists u_2 such that $f(u_2) \neq u_2 f(1)$. Obviously, $u_2 \neq 1$. Determinant $\det \mathbf{A}$ is 0 if and only if for all $u_1 \in \mathbb{R}$:

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

From nonlinearity of f again the equality above does not hold for a certain $u_1 \notin \{1, u_2\}$, otherwise we would have $\det \mathbf{A} = 0$. This ends the proof. □

Theorem 2.38. *For any uniformly continuous response function $q\colon \mathbb{R} \to (0,1)$ such that binary model is misspecified with respect to logistic loss (i.e. for all $a, b \in \mathbb{R}$ there exists $x \in \mathbb{R}:\ q(x) \neq q_L(ax+b)$) there exists $\mathbb{R}^{p+1}$-valued random variable $\boldsymbol{X} = (1, \tilde{\boldsymbol{X}}^T)^T$, for which $\tilde{\boldsymbol{X}}$ is supported on the set of non-zero Lebesgue measure and $s(\boldsymbol{X}) \cap s^*(\boldsymbol{X}) = \emptyset$.*

Proof. In order to prove the theorem we apply Lemma 2.36 to a discrete variable $\mathbf{Z}$ constructed as in Lemma 2.37 and $\beta_i(\mathbf{Z}) = \beta_i(\mathbf{Z}_m) = \beta_i = I(i \le k)$, $i = 1, \ldots, p$. Let $\mathbf{Z}_m$ from Lemma 2.36 for sufficiently large m be such that $s^*(\mathbf{Z}_m) = \{1, \ldots, p\}$. From the construction $s(\mathbf{Z}_m) = \{1, \ldots, k\}$. Let $X_i = Z_{mi}$ for $i \le k$, where Z_{mi} is defined in Lemma 2.36, $X_{k+1} = \sum_{i=1}^{k+1} \beta_i^*(\mathbf{Z}_m) Z_{mi}$, $X_{k+1+i} = \beta_{k+1+i}^*(\mathbf{Z}_m) Z_{m,k+1+i}$ for every $p - 1 - k \ge i > 0$. Then we show that $s(\mathbf{X}) = \{1, \ldots, k\}$, $s^*(\mathbf{X}) = \{k+1, \ldots, p\}$, that is $s(\mathbf{X}) \cap s^*(\mathbf{X}) = \emptyset$. Indeed, normal equations for the vector $\mathbf{Z}_m$ have the form

$$\mathbb{E} q_L(\boldsymbol{\beta}^{*T}(\mathbf{Z}_m)\mathbf{Z}_m)\mathbf{Z}_m = \mathbb{E} q(\boldsymbol{\beta}^T \mathbf{Z}_m)\mathbf{Z}_m = \mathbb{E} q\left(\sum_{i=1}^{k} \beta_i \mathbf{Z}_{mi}\right)\mathbf{Z}_m.$$

By rewriting them for vector $\mathbf{X}$, we obtain:

$$\mathbb{E} q_L\left(\sum_{i=k+1}^{p} X_i\right)\mathbf{X} = \mathbb{E} q\left(\sum_{i=1}^{k} \beta_i X_i\right)\mathbf{X}.$$

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

2.6. Sets of active predictors - examples

In this section we assume that function ρ defined in (2.2) is given by the formula:

$$\rho(b, y) = -by + \ln(1 + \exp(b)). \tag{2.40}$$

Example 2.39. *Let $X_t = \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t$ with $a_i \in \mathbb{R}, t \in \mathbb{Z}$ be a causal autoregressive AR(p) process where (ε_t) is a sequence of i.i.d. random variables with finite second moment (Brockwell and Davis, 1991, Chapter 3). Let $\tilde{\boldsymbol{X}}_1 = (X_n, \dots, X_1)^T$ and $\tilde{\boldsymbol{X}}_2 = X_{n+1}$ and $\boldsymbol{X} = (1, \tilde{\boldsymbol{X}}_1^T, \tilde{\boldsymbol{X}}_2^T)^T$ for $n \geq p$. Let $\mathbb{P}(Y = 1 | \boldsymbol{X} = (1, \tilde{\boldsymbol{x}}_1^T, \tilde{\boldsymbol{x}}_2^T)^T) = q(\beta_0 + \tilde{\boldsymbol{\beta}}_1^T \tilde{\boldsymbol{x}}_1)$ for some $(\beta_0, \tilde{\boldsymbol{\beta}}_1^T)^T \in \mathbb{R}^{n+1}$. We take $\boldsymbol{A} = (a_1, \dots, a_p, \boldsymbol{0}_{n-p}^T)$. Then from Theorem 2.2 we obtain $\boldsymbol{\beta}^* = (\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T}, 0)^T$, as $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2 - \boldsymbol{A}\tilde{\boldsymbol{X}}_1$ are independent.*

Example 2.40. *Let $p = 1$, $\boldsymbol{b} = (b_0, b_1)^T \in \mathbb{R}^2$,*

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}$$

and let $q(\boldsymbol{\beta}^T \boldsymbol{x}) = I(x_1 = 1)$. Then the risk function has the form:

$$\begin{aligned} R(\boldsymbol{b}) &= \mathbb{E} l(\boldsymbol{b}, \boldsymbol{X}, Y) = -\mathbb{E} Y(b_0 + b_1 X_1) + \mathbb{E} \ln\Big(1 + e^{b_0 + b_1 X_1}\Big) \\ &= -\frac{1}{2}(b_0 + b_1) + \frac{1}{2} \ln\Big(1 + e^{b_0 + b_1}\Big) + \frac{1}{2} \ln\Big(1 + e^{b_0 - b_1}\Big) \\ &= \frac{1}{2} \ln\Big((1 + e^{-b_0 - b_1})(1 + e^{b_0 - b_1})\Big). \end{aligned}$$

We observe that $\inf_{\boldsymbol{b} \in \mathbb{R}^2} R(\boldsymbol{b}) = 0$ (we take $\boldsymbol{b}_n = (0, n)^T$ for $n \in \mathbb{N}$), however R does not have global minimum. Thus $\boldsymbol{\beta}^*$ does not exist in this case. Moreover, assumption of Remark A.12 that $\mathbb{P}(Y = 1 | \boldsymbol{X} = \boldsymbol{x}) \in (0, 1)$ $\mathbb{P}_{\boldsymbol{X}}$ a.e. is not satisfied.

Example 2.41. *Let $\tilde{\boldsymbol{X}} \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} > 0$ and $\boldsymbol{X} = (1, \tilde{\boldsymbol{X}}^T)^T$. Consider a probit model for which $q(x) = \Phi(x)$ and $\boldsymbol{\beta} = (0, \tilde{\boldsymbol{\beta}}^T)^T$. To vector $(\boldsymbol{X}, Y)$ we fit logistic model. We have from symmetry of $\boldsymbol{X}$, q and q_L:*

$$\begin{aligned} 1 - \mathbb{E} q(\boldsymbol{\beta}^T \boldsymbol{X}) &= \mathbb{E} q(-\boldsymbol{\beta}^T \boldsymbol{X}) = \mathbb{E} q(\boldsymbol{\beta}^T \boldsymbol{X}) = \mathbb{E} q_L(\beta_0^* + \tilde{\boldsymbol{\beta}}^{*T} \tilde{\boldsymbol{X}}) \\ &= \mathbb{E} q_L(\beta_0^* - \tilde{\boldsymbol{\beta}}^{*T} \tilde{\boldsymbol{X}}) = 1 - \mathbb{E} q_L(-\beta_0^* + \tilde{\boldsymbol{\beta}}^{*T} \tilde{\boldsymbol{X}}), \end{aligned}$$

which implies:

$$\mathbb{E} q(\boldsymbol{\beta}^T \boldsymbol{X}) = \mathbb{E} q_L(\beta_0^* + \tilde{\boldsymbol{\beta}}^{*T} \tilde{\boldsymbol{X}}) = \mathbb{E} q_L(-\beta_0^* + \tilde{\boldsymbol{\beta}}^{*T} \tilde{\boldsymbol{X}}).$$

Hence $\beta_0^* = 0$ from uniqueness of $\boldsymbol{\beta}^*$. Let $U = \boldsymbol{\beta}^T \boldsymbol{X}$ and $\sigma^2 = \operatorname{Var} U = \tilde{\boldsymbol{\beta}}^T \boldsymbol{\Sigma} \tilde{\boldsymbol{\beta}}$. Thus $U \sim \mathcal{N}(0, \sigma^2)$. Then

$$\mathbb{E} q'(\boldsymbol{\beta}^T \boldsymbol{X}) = \mathbb{E} q'(U) = \int_{\mathbb{R}} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \cdot \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx = \frac{1}{\sqrt{2\pi}(\sigma^2 + 1)^{\frac{1}{2}}}.$$

This means that η in Theorem 2.6 satisfies the following inequality in view of Remark 2.24:

$$\eta \geq \frac{4}{\sqrt{2\pi(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}+1)}}. \tag{2.41}$$

We consider the previous example for general $\mathbf{X}$.

Example 2.42. *Let $\boldsymbol{X} = (1, \tilde{\boldsymbol{X}}^T)^T \in \mathbb{R}^{p+1}$ be a random vector with $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$ satisfying LRC($\boldsymbol{\beta}$) for $\boldsymbol{\beta} = (0, \tilde{\boldsymbol{\beta}}^T)^T \neq \boldsymbol{0}_{p+1}$, $\tilde{\boldsymbol{X}} \stackrel{d}{=} -\tilde{\boldsymbol{X}}$, LND and $q(x) = \Phi(x)$. Firstly, we observe that $\beta_0^* = 0$ as in Example 2.41 where only symmetry of $\tilde{\boldsymbol{X}}$ was used. Let $U = \tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}$. Left inequality in Theorem 4.1 in Alzer (2010) can be rewritten as:*

$$\Phi(x) > q_L\left(\sqrt{\frac{8}{\pi}}x\right)$$

for $x > 0$ and

$$\Phi(x) < q_L\left(\sqrt{\frac{8}{\pi}}x\right)$$

for $x < 0$. This means that for $x \neq 0$ we have:

$$\Phi(x)x > q_L\left(\sqrt{\frac{8}{\pi}}x\right)x.$$

Let $f(x) = \mathbb{E}q_L(xU)U$ for $x > 0$. Function f is strictly increasing as $q_L'(x) > 0$ for $x \in \mathbb{R}$ and we have $f'(x) = \mathbb{E}q_L'(xU)U^2 > 0$. In view of normal equations (2.10) we have:

$$f(\eta) = \mathbb{E}q_L(\eta U)U = \mathbb{E}\Phi(U)U > \mathbb{E}q_L\left(\sqrt{\frac{8}{\pi}}U\right)U = f\left(\sqrt{\frac{8}{\pi}}\right).$$

Hence

$$\eta > \sqrt{\frac{8}{\pi}}. \tag{2.42}$$

This lower bound is tighter than (2.41), as we have:

$$\sqrt{\frac{8}{\pi}} \geq \frac{4}{\sqrt{2\pi(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}+1)}}$$

and in contrast to (2.41) does not depend on $\boldsymbol{\beta}$ or $\boldsymbol{\Sigma}$.

Example below shows that we are able in case of some discrete distributions to give explicit formulas for $\boldsymbol{\beta}^*$ and moreover, that without LRC sets s and s^* can be disjoint (see also Examples 4.1 and 4.2 in Kubkowski and Mielniczuk (2017) for cases when $s^* \subset s$ and $s \subset s^*$ respectively).

Example 2.43. *Let vector $\tilde{\boldsymbol{X}} = (X_1, X_2)$ have the distribution:*

$$\mathbb{P}(X_1 = X_2 = 0) = \mathbb{P}(X_1 = X_2 = 1) = \mathbb{P}(X_1 = 7, X_2 = 5) = \frac{1}{3},$$

and

$$q(x) = \begin{cases} q_L(x) & \text{if } x \leq 1, \\ q_L(\frac{2}{3}x + \frac{1}{3}) & \text{otherwise.} \end{cases}$$

Let $\beta_0 = 0,\ \beta_1 = 1,\ \beta_2 = 0$. *The normal equations are:*

$$\mathbb{E}q_L(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2) = \mathbb{E}q(X_1),$$
$$\mathbb{E}q_L(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2)X_1 = \mathbb{E}q(X_1)X_1,$$
$$\mathbb{E}q_L(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2)X_2 = \mathbb{E}q(X_1)X_2,$$

which simplify to the form:

$$q_L(\beta_0^*) + q_L(\beta_0^* + \beta_1^* + \beta_2^*) + q_L(\beta_0^* + 7\beta_1^* + 5\beta_2^*) = q(0) + q(1) + q(7),$$
$$q_L(\beta_0^* + \beta_1^* + \beta_2^*) + 7q_L(\beta_0^* + 7\beta_1^* + 5\beta_2^*) = q(1) + 7q(7),$$
$$q_L(\beta_0^* + \beta_1^* + \beta_2^*) + 5q_L(\beta_0^* + 7\beta_1^* + 5\beta_2^*) = q(1) + 5q(7).$$

Hence after simple transformations we obtain:

$$q_L(\beta_0^*) = q(0),$$
$$q_L(\beta_0^* + 7\beta_1^* + 5\beta_2^*) = q(7),$$
$$q_L(\beta_0^* + \beta_1^* + \beta_2^*) = q(1).$$

Similarly as before we have:

$$\beta_0^* = 0,$$
$$\beta_0^* + 7\beta_1^* + 5\beta_2^* = 5,$$
$$\beta_0^* + \beta_1^* + \beta_2^* = 1.$$

Hence $\beta_2^* = 1, \beta_1^* = \beta_0^* = 0$. *This means that sets* $s^* = \{2\}$ *and* $s = \{1\}$ *are disjoint:* $s^* \cap s = \emptyset$.

We show now an example of $s \cap s^* = \emptyset$ for continuous $\tilde{\mathbf{X}}$.

Example 2.44. *Let* $q(x) = q(-x)$, X_1, ε *are independent and* $\mathcal{U}[-1,1]$ *distributed,* $X_2 = k(X_1 + l\varepsilon)^2$ *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 we have* $\beta_2^* \neq 0$.
Firstly, let us observe that:

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

Let $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \beta_2^*)^T$ *be corresponding projection of* $\beta_1 = 1,\ \beta_2 = \beta_0 = 0$ *in fitted logistic model and* $\bar{\boldsymbol{\beta}} = (\bar{\beta}_0, \bar{\beta}_1, \bar{\beta}_2)^T$ *be 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* $\bar{\beta}_0 = \beta_0^*, \bar{\beta}_1 = -\beta_1^*, \bar{\beta}_2 = \beta_2^*$. *On the other hand, symmetry of* q *implies that* $q(X_1) = q(-X_1)$, *hence we have* $\boldsymbol{\beta}^* = \bar{\boldsymbol{\beta}}$ *from uniqueness of projection. This means that* $\beta_1^* = 0$.
Suppose now that $\beta_2^* = 0$. *Normal equations take the form:*

$$\mathbb{E}q(X_1) = q_L(\beta_0^*),$$
$$\mathbb{E}q(X_1)X_1 = q_L(\beta_0^*)\mathbb{E}X_1,$$

$$\mathbb{E}q(X_1)X_2 = q_L(\beta_0^*)\mathbb{E}X_2.$$

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

$$\mathbb{E}q(X_1)X_2 = q_L(\beta_0^*)\mathbb{E}X_2 = \mathbb{E}q(X_1)\mathbb{E}X_2.$$

This means that $\mathrm{Cov}(q(X_1), X_2) = 0$, *contradicting the assumptions, thus* $\beta_2^* \neq 0$.
Figure 2.1 shows direction of ML estimate from logistic model when $k = 2, l = 0.25$ *and* $q(t) = 0.75 - 0.5 \cdot t^2$ *for* $t \in [-1, 1]$.

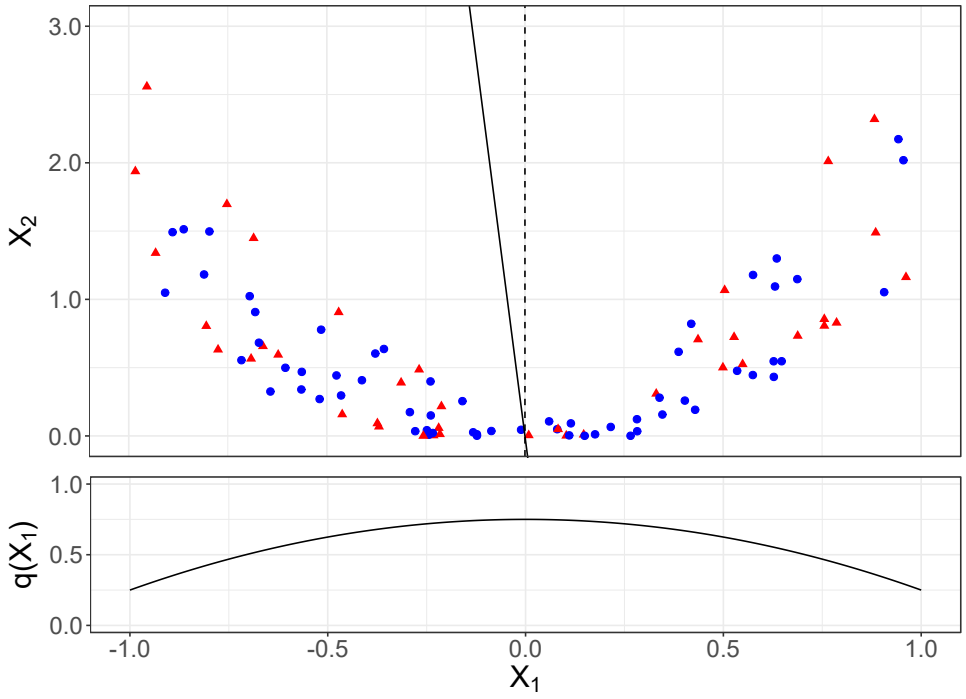


Figure 2.1: Scatter plot pertaining to the distribution in Example 2.44. Triangles and circles correspond to $Y = 0$ and $Y = 1$, respectively. Solid line shows the direction of estimator $\hat{\boldsymbol{\beta}}$ based on fitted logistic model. The form of q is depicted in the lower plot.

Example 2.45. *Let* $q(x) = q_L(x^3)$, $\boldsymbol{X} = (1, \tilde{\boldsymbol{X}}^T)^T$, *where* $\tilde{\boldsymbol{X}} = (X_1, \ldots, X_p)^T \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ *and* $\boldsymbol{\Sigma} = [\rho^{|i-j|}]_{1 \leq i,j \leq p}$. *Let* $\mathbb{P}(Y = 1|\boldsymbol{X}) = q(X_1 + X_2) = q_L((X_1 + X_2)^3)$. *This model is misspecified and vector* $\boldsymbol{X}$ *satisfies linear regressions condition. Hence from the Theorem 2.16 it follows that* $\tilde{\boldsymbol{\beta}}^* = \eta\tilde{\boldsymbol{\beta}} = (\eta, \eta, \boldsymbol{0}_{p-2}^T)^T$.
We see that $s = \{1, 2\}$. *Now we will prove that* $\eta > 0$ *and* $s^* = s$. *To see this observe that function* q *is strictly increasing. From Lemma A.44 we have that:*

$$0 < \mathrm{Cov}(\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}, q(\boldsymbol{\beta}^T\boldsymbol{X})) = \tilde{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}a_{\boldsymbol{\beta}} = ||\boldsymbol{\Sigma}^{\frac{1}{2}}\tilde{\boldsymbol{\beta}}||_2^2 a_{\boldsymbol{\beta}}.$$

Hence $a_{\boldsymbol{\beta}} > 0$, *moreover from Remark 2.17 we obtain* $a_{\boldsymbol{\beta}^*} > 0$. *Thus* $\eta = \frac{a_{\boldsymbol{\beta}}}{a_{\boldsymbol{\beta}^*}} > 0$ *and* $s^* = s$.

Example 2.46. *Let* $q(x) = q_L(x^3)$, $\boldsymbol{Z} = (1, \tilde{\boldsymbol{Z}}^T)^T$, *where* $\tilde{\boldsymbol{Z}} = (Z_1, \ldots, Z_p)^T \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ *and* $\boldsymbol{\Sigma} = [\rho^{|i-j|}]_{1 \leq i,j \leq p}$. *Let* $\boldsymbol{X} = (1, X_1, \ldots, X_p)^T$, *where* $X_1 = Z_1, X_2 = Z_2, X_3 = Z_1^3, X_4 = Z_2^3, X_5 = Z_1^2 Z_2, X_6 = Z_1 Z_2^2, X_7 = Z_1^2, X_8 = Z_2^2, X_9 = Z_1 Z_2$ *and* $X_{i+7} = Z_i$ *for* $i = 3, \ldots, p$. *Let*

$$\mathbb{P}(Y = 1|\boldsymbol{X}) = q(X_1 + X_2) = q_L((X_1 + X_2)^3) = q_L(X_3 + X_4 + 3X_5 + 3X_6).$$

This model is clearly well specified. This means that $s^* = \{3, 4, 5, 6\}$. *Analogously, as in the Example 2.45 we obtain* $s = \{1, 2\}$. *This means that* $s \cap s^* = \emptyset$. *Note that vector* $\boldsymbol{X}$ *does not satisfy linear regressions condition in this example. The above model (with different ordering of predictors) will be considered in numerical experiments.*

Chapter 3

Properties of the projection in the generalized semiparametric model

In this chapter we consider generalized semiparametric binary model:

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q(\mathbf{x}^T\boldsymbol{\beta}_1, \ldots, \mathbf{x}^T\boldsymbol{\beta}_k) = q(\mathbf{B}^T\mathbf{X}), \tag{3.1}$$

where $k \in \mathbb{N}$, $\mathbf{x}, \boldsymbol{\beta}_i \in \Theta = \mathbb{R}^{p_n+1}$ for $i = 1, \ldots, k$. Analogously to the Chapter 2, we assume that $q^{(n)} = q : \mathbb{R}^k \to [0, 1]$, $p_n = p \in \mathbb{N}$. We additionally assume that $k \leq p$ and $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k] \in \mathbb{R}^{(p+1)\times k}$. Because our focus will be on non-constant predictors and not on the intercept, we introduce the following notation: $\mathbf{X} = (X_0, \tilde{\mathbf{X}}^T)^T$, $\tilde{\mathbf{X}} = (X_1, \ldots, X_p)^T$, $X_0 \equiv 1$, $\tilde{\mathbf{b}} = (b_1, \ldots, b_p)^T, \mathbf{b} = (b_0, \tilde{\mathbf{b}}^T)^T, \tilde{\mathbf{B}} = [\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_k]$, $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k] = [\mathbf{B}_0^T, \tilde{\mathbf{B}}^T]^T$, where $\boldsymbol{\beta}_i = (\beta_{i0}, \tilde{\boldsymbol{\beta}}_i^T)^T$ and $\mathbf{B}_0 = [\beta_{10}, \ldots, \beta_{k0}]$. If the appropriate moments of $\tilde{\mathbf{X}}$ are finite, we write $\mathbb{E}\tilde{\mathbf{X}} = \boldsymbol{\mu}$, $\operatorname{Var}\tilde{\mathbf{X}} = \boldsymbol{\Sigma}$.

Our main aim in this chapter is to show how to extend results from the Chapter 2 to the case of generalized semiparametric binary model with the focus on results related to logistic loss (see Sections 3.1-3.2). Moreover, in Section 3.4 we consider additive binary model, which is a special case of the model considered in this chapter. Section 3.3 presents a result interesting in its own right, which shows that direction obtained by LDA method is the same as direction of $\tilde{\boldsymbol{\beta}}^*$ under linear regressions condition (see Theorem 3.20).

Linear regressions condition has more general form in this chapter than in Chapter 2 (matrix $\mathbf{C}$ has analogous structure to matrix $\mathbf{B}$):

$$\text{(LRC}(\mathbf{C})) \;\; \exists \mathbf{h}_0 = \mathbf{h}_0(\mathbf{C}) \in \mathbb{R}^p, \mathbf{H} = \mathbf{H}(\mathbf{C}) \in \mathbb{R}^{p\times k} : \; \mathbb{E}(\tilde{\mathbf{X}}|\tilde{\mathbf{C}}^T\tilde{\mathbf{X}}) = \mathbf{h}_0 + \mathbf{H}\tilde{\mathbf{C}}^T\tilde{\mathbf{X}},$$

Condition LRC($\mathbf{C}$) is satisfied for every $\mathbf{C} \in \mathbb{R}^{(p+1)\times k}$ with $\operatorname{rank}\tilde{\mathbf{C}} = k$, where $\tilde{\mathbf{X}}$ has elliptically contoured distributions with finite second moment (see Section A.2). Moreover, if $k < p$ and the condition LRC($\mathbf{C}$) is satisfied for every $\mathbf{C} \in \mathbb{R}^{(p+1)\times k}$ with $\operatorname{rank}\tilde{\mathbf{C}} = k$, then $\tilde{\mathbf{X}}$ has elliptically contoured distribution (see Theorem A.23). Note that normal distribution belongs to the family of elliptically contoured distributions (see Remark A.17).

We will show in the Theorem 3.4 that the condition LRC($\mathbf{B}$) is essential in the proof of equality $\tilde{\boldsymbol{\beta}}^* = \tilde{\mathbf{B}}\boldsymbol{\eta}$ for some $\boldsymbol{\eta} \in \mathbb{R}^k$. Condition LRC($\boldsymbol{\beta}^*$) allow us to represent $\boldsymbol{\beta}^*$ as some function of $\mathbf{B}$ even in the situation when some predictors are omitted in the fitted model - in the case of logistic loss see Section 3.2.
In this chapter we assume that loss function is of the form:

$$l(\mathbf{b}, \mathbf{x}, y) = \rho(\mathbf{b}^T\mathbf{x}, y), \tag{3.2}$$

where $\rho : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ is some function, $\mathbf{b}, \mathbf{x} \in \mathbb{R}^{p+1}, y \in \{0, 1\}$. In almost all theorems (except Theorem 3.1) we assume that $\rho(\cdot, y)$ is convex (or strictly convex) function for all y. Differentiability of $\rho(\cdot, y)$ for all y along with LRC is used in Sections 3.1, 3.2 and 3.4 to show the form of $\boldsymbol{\beta}^*$ in the fitted model.

We will also consider LND condition introduced in Chapter 2. We note that discussion before Section 2.1 remains valid for generalized semiparametric model.

The results discussed below are slight modifications of Kubkowski and Mielniczuk (2018).

3.1. General loss

The following theorem is a generalization of Theorem 2.2. It states that when inactive predictors $\tilde{\mathbf{X}}_2$ in binary model are such that $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2 - \mathbf{A}\tilde{\mathbf{X}}_1$ are independent, where $\tilde{\mathbf{X}}_1$ are remaining predictors and $\mathbf{A}$ is a linear transform, then minimizer $\boldsymbol{\beta}^*$ of the risk in the model containing $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ is obtained from the minimizer $(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T})$ of the risk in the model containing only $\tilde{\mathbf{X}}_1$ by appending zeroes to the latter. It is easily seen that the result fails if $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ are dependent (see Example 2.43).

Theorem 3.1. *Let $\boldsymbol{X}$ be a random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$ and assume that $\mathbb{E}|\rho(\boldsymbol{b}^T\boldsymbol{X}, Y)| < \infty$ for all $\boldsymbol{b} \in \mathbb{R}^{p+1}$. Let $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}_1^T, \tilde{\boldsymbol{X}}_2^T)^T$, where $\tilde{\boldsymbol{X}}_1^T = (X_1, \ldots, X_j)^T$, $\tilde{\boldsymbol{X}}_2^T = (X_{j+1}, \ldots, X_p)^T$, $\tilde{\boldsymbol{B}} = (\tilde{\boldsymbol{B}}_1^T, \tilde{\boldsymbol{B}}_2^T)^T$, where $\tilde{\boldsymbol{B}}_1 \in \mathbb{R}^{j\times k}, \tilde{\boldsymbol{B}}_2 \in \mathbb{R}^{(p-j)\times k}$. Assume that $\tilde{\boldsymbol{B}}_2 = \boldsymbol{O}_{(p-j)\times k}$ is matrix with elements equal 0 and that $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2 - \boldsymbol{A}\tilde{\boldsymbol{X}}_1$ are independent for a certain $\boldsymbol{A} \in \mathbb{R}^{(p-j)\times j}$. If there exists $(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T})^T$ such that:*

$$(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T})^T = \underset{(b_0, \boldsymbol{b}_1^T)^T \in \mathbb{R}^{j+1}}{\arg\min} \mathbb{E}\rho(b_0 + \boldsymbol{b}_1^T\tilde{\boldsymbol{X}}_1, Y), \tag{3.3}$$

then $\boldsymbol{\beta}^*$ defined in (1.3) exists and:

$$\boldsymbol{\beta}^* = (\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T}, \boldsymbol{0}_{p-j}^T)^T.$$

Moreover, if we assume LND and strict convexity of $\rho(\cdot, y)$ for all y, then $\boldsymbol{\beta}^*$ is unique.

Proof. Let (analogously as in the Theorem 2.2):

$$h(b_0, \mathbf{b}_1^T, \mathbf{b}_2^T) = \mathbb{E}\rho(b_0 + \mathbf{b}_1^T\tilde{\mathbf{X}}_1 + \mathbf{b}_2^T\tilde{\mathbf{X}}_2, Y).$$

The proof is now identical to the Theorem 2.2 (we need only to replace $\beta_0, \tilde{\boldsymbol{\beta}}_1$ by $\mathbf{B}_0, \tilde{\mathbf{B}}_1$, respectively). □

Remark 3.2. *If we fit the model without intercept (see Definition 1.8) in the Theorem 3.1, we need additionally to assume $\mathbb{E}\tilde{\boldsymbol{X}} = \boldsymbol{0}_p$ for Theorem 3.1 to remain valid and the proof is analogous as in the case of semiparametric setup.*

Remark 3.3. *Note that Theorem 3.1 holds in particular when $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$ are independent.*

The following theorem is a generalization of Theorem 1 in Kubkowski and Mielniczuk (2018) to the case of any convex loss. Moreover, it can be also viewed as generalization of Theorem 2.6 to the case of generalized semiparametric model, as the proof is similar.

Theorem 3.4. *Let $\boldsymbol{X}$ be a random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $\rho(\cdot, y)$ is convex function for all y, condition*

$$\forall \boldsymbol{b} \in \mathbb{R}^{p+1} : \ \mathbb{E}|\rho(\boldsymbol{b}^T \boldsymbol{X}, Y)| < \infty,$$

is satisfied and assume LRC($\boldsymbol{B}$). If there exists $\beta_0^* \in \mathbb{R}, \boldsymbol{\eta} \in \mathbb{R}^k$ such that:

$$(\beta_0^*, \boldsymbol{\eta}^T)^T = \underset{(b_0, \boldsymbol{c}^T)^T \in \mathbb{R}\times\mathbb{R}^k}{\arg\min} \ \mathbb{E}\rho(b_0 + \boldsymbol{c}^T \tilde{\boldsymbol{B}}^T \tilde{\boldsymbol{X}}), \tag{3.4}$$

then $\boldsymbol{\beta}^*$ defined in (1.3) exists and is a linear combination of $\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_k$:

$$\tilde{\boldsymbol{\beta}}^* = \sum_{i=1}^{k} \eta_i \tilde{\boldsymbol{\beta}}_i = \tilde{\boldsymbol{B}}\boldsymbol{\eta}. \tag{3.5}$$

Moreover, if we assume LND and strict convexity of $\rho(\cdot, y)$ for all y, then $\boldsymbol{\beta}^*$ is unique.

Proof. Let $\mathbf{r} \in \mathbb{R}^p$, $\mathbf{c} \in \mathbb{R}^k$ and $\tilde{\mathbf{b}} = \tilde{\mathbf{B}}\mathbf{c} + \mathbf{r}$. Then loss l can be written as:

$$l(\mathbf{b}, \mathbf{X}, Y) = \rho(b_0 + \mathbf{c}^T \tilde{\mathbf{B}}^T \tilde{\mathbf{X}} + \mathbf{r}^T \tilde{\mathbf{X}}, Y) =: h(b_0, \mathbf{c}, \mathbf{r}).$$

We define function $J(b_0, \mathbf{c}, \mathbf{r}) = \mathbb{E}h(b_0, \mathbf{c}, \mathbf{r})$. J is well defined in view of moment assumptions about ρ. We observe that (3.4) is equivalent to: $J(\beta_0^*, \boldsymbol{\eta}, \mathbf{0}_p) \leq J(b_0, \mathbf{c}, \mathbf{0}_p)$ for every $b_0 \in \mathbb{R}$ and $\mathbf{c} \in \mathbb{R}^k$. For the first part of the theorem, we need only to show that

$$(\beta_0^*, \boldsymbol{\eta}, \mathbf{0}_p) = \underset{(b_0, \mathbf{c}, \mathbf{r}) \in \mathbb{R}\times\mathbb{R}^k\times\mathbb{R}^p}{\arg\min} \ J(b_0, \mathbf{c}, \mathbf{r}). \tag{3.6}$$

Now, by conditioning on $\tilde{\mathbf{X}}$ and then on $\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}$, from LRC($\mathbf{B}$), Jensen's inequality and (3.4) we obtain:

$$\begin{aligned}
J(b_0, \mathbf{c}, \mathbf{r}) &= \mathbb{E}\rho(b_0 + \mathbf{c}^T \tilde{\mathbf{B}}^T \tilde{\mathbf{X}} + \mathbf{r}^T \tilde{\mathbf{X}}, Y) \\
&= \mathbb{E}(\mathbb{E}(\rho(b_0 + \mathbf{c}^T \tilde{\mathbf{B}}^T \tilde{\mathbf{X}} + \mathbf{r}^T \tilde{\mathbf{X}}, Y)|\tilde{\mathbf{X}})) \\
&= \mathbb{E}\rho(b_0 + \mathbf{c}^T \tilde{\mathbf{B}}^T \tilde{\mathbf{X}} + \mathbf{r}^T \tilde{\mathbf{X}}, 1) q(\mathbf{B}^T\mathbf{X}) \\
&\quad + \mathbb{E}\rho(b_0 + \mathbf{c}^T \tilde{\mathbf{B}}^T \tilde{\mathbf{X}} + \mathbf{r}^T \tilde{\mathbf{X}}, 0)(1 - q(\mathbf{B}^T\mathbf{X})) \\
&= \mathbb{E}(\mathbb{E}(\rho(b_0 + \mathbf{c}^T \tilde{\mathbf{B}}^T \tilde{\mathbf{X}} + \mathbf{r}^T \tilde{\mathbf{X}}, 1)|\tilde{\mathbf{B}}^T \tilde{\mathbf{X}}) q(\mathbf{B}^T\mathbf{X})) \\
&\quad + \mathbb{E}(\mathbb{E}(\rho(b_0 + \mathbf{c}^T \tilde{\mathbf{B}}^T \tilde{\mathbf{X}} + \mathbf{r}^T \tilde{\mathbf{X}}, 0)|\tilde{\mathbf{B}}^T \tilde{\mathbf{X}})(1 - q(\mathbf{B}^T\mathbf{X})))
\end{aligned}$$

$$
\begin{aligned}
&\geq \mathbb{E}\rho(\mathbb{E}(b_0 + \mathbf{c}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}), 1)q(\mathbf{B}^T\mathbf{X}) \\
&\quad + \mathbb{E}\rho(\mathbb{E}(b_0 + \mathbf{c}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}), 0)(1 - q(\mathbf{B}^T\mathbf{X})) \\
&= \mathbb{E}\rho(\mathbb{E}(b_0 + \mathbf{c}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}} + \mathbf{r}^T\tilde{\mathbf{X}}|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}), Y) \\
&= \mathbb{E}\rho(b_0 + \mathbf{c}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}} + \mathbf{r}^T(\mathbf{h}_0 + \mathbf{H}\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}), Y) \\
&= J(b_0 + \mathbf{r}^T\mathbf{h}_0, \mathbf{c} + \mathbf{H}^T\mathbf{r}, \mathbf{0}_p) \geq J(\beta_0^*, \boldsymbol{\eta}, \mathbf{0}_p).
\end{aligned}
$$

This means that point $(\beta_0^*, \boldsymbol{\eta}, \mathbf{0}_p)$ is a global minimum of J. Hence (3.6) is satisfied and $\boldsymbol{\beta}^* = (\beta_0^*, \tilde{\mathbf{B}}^T\boldsymbol{\eta})^T$ exists. Uniqueness of $\boldsymbol{\beta}^*$ is obtained by using similar reasoning as in the proof of the Theorem 2.2. □

Corollary 3.5. *If l is logistic loss, $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $q(\boldsymbol{B}^T\boldsymbol{X}) \in (0,1)$ $\mathbb{P}_{\boldsymbol{X}}$ a.e., LND and LRC($\boldsymbol{B}$) hold then $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\eta}^T\tilde{\boldsymbol{B}}^T)^T$ for some $\beta_0^* \in \mathbb{R}, \boldsymbol{\eta} \in \mathbb{R}^k$.*

Proof. From Remark A.12 we obtain that solution of (3.4) exists and moment assumptions of Theorem 3.4 are satisfied. This ends the proof. □

Analogously, for quadratic and probit losses we obtain in view of Theorem 3.4, note above Theorem 2.6 and Remarks A.13-A.14:

Corollary 3.6. *If l is quadratic loss, $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$, $\boldsymbol{\Sigma} > 0$ and LRC($\boldsymbol{B}$) holds then $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\eta}^T\tilde{\boldsymbol{B}}^T)^T$ for some $\beta_0^* \in \mathbb{R}, \boldsymbol{\eta} \in \mathbb{R}^k$.*

Corollary 3.7. *If l is probit loss, $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$, $\boldsymbol{\Sigma} > 0$, $q(\boldsymbol{B}^T\boldsymbol{X}) \in (0,1)$ $\mathbb{P}_{\boldsymbol{X}}$ a.e. and LRC($\boldsymbol{B}$) holds then $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\eta}^T\tilde{\boldsymbol{B}}^T)^T$ for some $\beta_0^* \in \mathbb{R}, \boldsymbol{\eta} \in \mathbb{R}^k$.*

In the case of the model without intercept, Theorem 3.4 has the following form (with additional assumption $\mathbb{E}\tilde{\mathbf{X}} = \mathbf{0}_p$) and the proof is analogous:

Theorem 3.8. *Let $\boldsymbol{X}$ be a random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $\mathbb{E}\tilde{\boldsymbol{X}} = \boldsymbol{0}_p$, $\rho(\cdot, y)$ is convex function for all y, condition*

$$\forall \boldsymbol{b} \in \mathbb{R}^{p+1}\colon\ \mathbb{E}|\rho(\boldsymbol{b}^T\boldsymbol{X}, Y)| < \infty,$$

is satisfied and assume LRC($\boldsymbol{B}$). If there exists $\boldsymbol{\eta} \in \mathbb{R}^k$ such that:

$$\boldsymbol{\eta} = \arg\min_{\boldsymbol{c} \in \mathbb{R}^k} \mathbb{E}\rho(\boldsymbol{c}^T\tilde{\boldsymbol{B}}^T\boldsymbol{X}, Y),$$

then $\boldsymbol{\beta}^*$ defined in (1.3) exists and:

$$\boldsymbol{\beta}^* = \tilde{\boldsymbol{B}}\boldsymbol{\eta}.$$

Moreover, if we assume LND and strict convexity of $\rho(\cdot, y)$ for all y, then $\boldsymbol{\beta}^*$ is unique.

Remark 3.9. *Note that when $\rho(\cdot, y)$ is differentiable for all y,*

$$\forall \boldsymbol{b} \in \mathbb{R}^{p+1}\colon\ \mathbb{E}|\rho(\boldsymbol{b}^T\boldsymbol{X}, Y)| < \infty,$$
$$\exists h : \mathbb{R}^{p+1} \times \{0,1\} \to \mathbb{R}\ \forall \boldsymbol{b} \in \mathbb{R}^{p+1}\colon\ \left|\left|\frac{\partial \rho}{\partial b}(\boldsymbol{b}^T\boldsymbol{X}, Y)\boldsymbol{X}\right|\right|_2 \leq h(\boldsymbol{X}, Y)$$

where $\mathbb{E}h(\boldsymbol{X}, Y) < \infty$ *and* $\boldsymbol{\beta}^*$ *exists, then* $\boldsymbol{\beta}^*$ *is solution to normal equations (the same as in the semiparametric setup - see (2.9)):*

$$\mathbb{E}\left(\frac{\partial \rho}{\partial b}(\boldsymbol{b}^T\boldsymbol{X}, Y)\boldsymbol{X}\right) = 0. \tag{3.7}$$

Furthermore, if $\rho(\cdot, y)$ *is convex function for all* y *and (3.7) is satisfied for* $\boldsymbol{b} = \boldsymbol{\beta}^*$*, then* $\boldsymbol{\beta}^*$ *is a minimizer of risk function* R*.*

In the case of logistic loss, expression (3.7) reduces further to normal equations for logistic regression which are the same as in the semiparametric setup (compare with (2.10)):

$$\mathbb{E}(Y - q_L(\boldsymbol{\beta}^{*T}\mathbf{X}))\mathbf{X} = 0. \tag{3.8}$$

By conditioning the above expectation on $\mathbf{X}$, we obtain analogous equations as in semiparametric setup:

$$\mathbb{E}(q(\mathbf{B}^T\mathbf{X}) - q_L(\boldsymbol{\beta}^{*T}\mathbf{X}))\mathbf{X} = 0. \tag{3.9}$$

In the model with intercept we obtain additionally ($\mathbb{E}Y = \mathbb{E}q_L(\boldsymbol{\beta}^{*T}\mathbf{X})$):

$$\operatorname{Cov}(\tilde{\mathbf{X}}, Y) = \operatorname{Cov}(\tilde{\mathbf{X}}, q(\mathbf{B}^T\mathbf{X})) = \operatorname{Cov}(\tilde{\mathbf{X}}, q_L(\boldsymbol{\beta}^{*T}\mathbf{X})). \tag{3.10}$$

Similar equations can be written down in the case of probit, quadratic and Huber losses analogously to (2.13)-(2.17) with $\mathbf{B}$ instead of $\boldsymbol{\beta}$.

The following lemma is a generalization of Lemma 2.14, which was proved in the case of the semiparametric setup.

Lemma 3.10. *Let* $\boldsymbol{X}$ *be random vector such that* $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$ *with* $\mathbb{E}\tilde{\boldsymbol{X}} = \boldsymbol{\mu}$ *and* $\operatorname{Var}\tilde{\boldsymbol{X}} = \boldsymbol{\Sigma} > 0$*. If* $\boldsymbol{X}$ *satisfies LRC(*$\boldsymbol{B}$*), where* $\operatorname{rank}\tilde{\boldsymbol{B}} = k$ *then we have:*

$$\boldsymbol{H} = \boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(\tilde{\boldsymbol{B}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{B}})^{-1}, \tag{3.11}$$

$$\boldsymbol{h}_0 = (\boldsymbol{I}_p - \boldsymbol{H}\tilde{\boldsymbol{B}}^T)\boldsymbol{\mu} = (\boldsymbol{I}_p - \boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(\tilde{\boldsymbol{B}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{B}}^T)\boldsymbol{\mu}. \tag{3.12}$$

Proof. Observe that

$$\operatorname{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) = \operatorname{Cov}(\mathbb{E}(\tilde{\mathbf{X}}|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}), \mathbb{E}(\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}})) + \mathbb{E}\operatorname{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) \tag{3.13}$$

$$= \operatorname{Cov}(\mathbb{E}(\tilde{\mathbf{X}}|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}), \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) + \mathbf{0}_p = \operatorname{Cov}(\mathbf{h}_0 + \mathbf{H}\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}, \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}). \tag{3.14}$$

As $\mathbf{h}_0$ and $\mathbf{H}$ are deterministic, we have

$$\operatorname{Cov}(\mathbf{h}_0 + \mathbf{H}\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}, \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) = \mathbf{H}\operatorname{Var}(\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) = \mathbf{H}\tilde{\mathbf{B}}^T\boldsymbol{\Sigma}\tilde{\mathbf{B}}. \tag{3.15}$$

Because $\operatorname{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) = \boldsymbol{\Sigma}\tilde{\mathbf{B}}$ it follows from Equations (3.13) and (3.15) that

$$\boldsymbol{\Sigma}\tilde{\mathbf{B}} = \mathbf{H}\tilde{\mathbf{B}}^T\boldsymbol{\Sigma}\tilde{\mathbf{B}}$$

and thus $\mathbf{H} = \boldsymbol{\Sigma}\tilde{\mathbf{B}}(\tilde{\mathbf{B}}^T\boldsymbol{\Sigma}\tilde{\mathbf{B}})^{-1}$. Formula for $\mathbf{h}_0$ follows after taking expected values of both sides of LRC($\mathbf{B}$) and using formula for $\mathbf{H}$. □

Remark 3.11. *Note that as* $\tilde{\boldsymbol{B}}^T\boldsymbol{H} = \boldsymbol{I}_p$ *it follows from Lemma 3.10 that* $\boldsymbol{\beta}_i$ *is perpendicular to* $\boldsymbol{H}^{(j)}$ *for* $i \neq j$*. Moreover, we have that* $\boldsymbol{h}_0$ *is perpendicular to all* $\boldsymbol{\beta}_j$*, as* $\boldsymbol{B}^T\boldsymbol{h}_0 = (\boldsymbol{B}^T - \boldsymbol{B}^T\boldsymbol{H}\boldsymbol{B}^T)\boldsymbol{\mu} = \boldsymbol{0}_k$*.*

3.2. Logistic loss

In this section we assume that function ρ defined in (2.2) is given by the formula:

$$\rho(b, y) = -by + \ln(1 + \exp(b)). \tag{3.16}$$

Here we give another proof of Theorem 3.4 for logistic loss based on normal equations (3.8). The following theorem gives sufficient condition when the vector $\boldsymbol{\eta}$ is nonzero. Method of the proof is based on Brillinger (1982) where the proportionality constant was obtained for linear model. This method allows us to represent $\boldsymbol{\beta}^*$ as a function of $\mathbf{B}$ even in the situation where some predictors are omitted what is shown in Proposition 3.17. We define:

$$a_{\mathbf{B}} = (\operatorname{Var}(\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}))^{-1}\operatorname{Cov}(\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}, Y) \in \mathbb{R}^k. \tag{3.17}$$

Theorem 3.12. *Let $\boldsymbol{X}$ be random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$ and $\boldsymbol{\Sigma} > 0$. If $\operatorname{rank}\tilde{\boldsymbol{B}} = k$, $\boldsymbol{X}$ satisfies LRC($\boldsymbol{B}$) and LRC($\boldsymbol{\beta}^*$) then we have*

$$\tilde{\boldsymbol{\beta}}^* a_{\beta^*} = \tilde{\boldsymbol{B}} a_{\boldsymbol{B}}, \tag{3.18}$$

where

$$a_{\boldsymbol{\beta}^*} = (\operatorname{Var}(\tilde{\boldsymbol{\beta}}^{*T}\tilde{\boldsymbol{X}}))^{-1}\operatorname{Cov}(\tilde{\boldsymbol{\beta}}^{*T}\tilde{\boldsymbol{X}}, Y).$$

Moreover if $\operatorname{Cov}(\boldsymbol{B}^T\boldsymbol{X}, Y) \neq \boldsymbol{0}_k$, then $a_{\boldsymbol{B}}, a_{\beta^*}$ and $\tilde{\boldsymbol{\beta}}^*$ are nonzero.

Proof. Using covariance decomposition with conditioning vector $\mathbf{B}^T\mathbf{X}$ and Theorem 3.10, we obtain

$$\begin{aligned}\operatorname{Cov}(\tilde{\mathbf{X}}, q(\mathbf{B}^T\mathbf{X})) &= \operatorname{Cov}(\mathbb{E}(\tilde{\mathbf{X}}|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}), q(\mathbf{B}^T\mathbf{X})) + \mathbb{E}\operatorname{Cov}(\tilde{\mathbf{X}}, q(\mathbf{B}^T\mathbf{X})|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) = \\ &= \operatorname{Cov}(\mathbf{h}_0 + \mathbf{H}\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}, q(\mathbf{B}^T\mathbf{X})) = \\ &= \boldsymbol{\Sigma}\tilde{\mathbf{B}}(\tilde{\mathbf{B}}^T\boldsymbol{\Sigma}\tilde{\mathbf{B}})^{-1}\operatorname{Cov}(\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}, q(\mathbf{B}^T\mathbf{X})) = \boldsymbol{\Sigma}\tilde{\mathbf{B}}a_{\mathbf{B}},\end{aligned} \tag{3.19}$$

as $\operatorname{Cov}(\tilde{\mathbf{X}}, q(\mathbf{B}^T\mathbf{X})|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) = \mathbf{0}_p$ and the last equality follows from the definition of $a_{\mathbf{B}}$. Analogously, using linear regressions condition for $\tilde{\boldsymbol{\beta}}^*$ and Theorem 3.10 for $k = 1$ and $\tilde{\boldsymbol{\beta}}^*$ we obtain the equality:

$$\operatorname{Cov}(\tilde{\mathbf{X}}, q_L(\beta_0^* + \tilde{\mathbf{X}}^T\tilde{\boldsymbol{\beta}}^*)) = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}^* a_{\beta^*}. \tag{3.20}$$

From the normal equations (3.10) we have:

$$\operatorname{Cov}(\tilde{\mathbf{X}}, q_L(\beta_0^* + \tilde{\mathbf{X}}^T\tilde{\boldsymbol{\beta}}^*)) = \operatorname{Cov}(\tilde{\mathbf{X}}, q(\mathbf{B}^T\mathbf{X})).$$

Thus from (3.19) and (3.20) it follows:

$$\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}^* a_{\beta^*} = \boldsymbol{\Sigma}\tilde{\mathbf{B}}a_{\mathbf{B}}.$$

From the invertibility of the matrix $\boldsymbol{\Sigma}$ the first part of the Theorem follows. To prove the second part, we observe that:

$$\mathbf{0}_k \neq \operatorname{Cov}(\mathbf{B}^T\mathbf{X}, Y) = \operatorname{Cov}(\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}, q(\mathbf{B}^T\mathbf{X})) = \tilde{\mathbf{B}}^T\boldsymbol{\Sigma}\tilde{\mathbf{B}}a_{\mathbf{B}}.$$

Because $\mathbf{\Sigma} > 0$ and $\operatorname{rank} \tilde{\mathbf{B}} = k$, hence $\tilde{\mathbf{B}}^T \mathbf{\Sigma} \tilde{\mathbf{B}} > 0$. This means that $a_{\mathbf{B}} \neq \mathbf{0}_k$. From the first part of the Theorem and again from assumption $\operatorname{rank} \tilde{\mathbf{B}} = k$ we have $\tilde{\mathbf{B}} a_{\mathbf{B}} \neq \mathbf{0}_p$ and $\tilde{\boldsymbol{\beta}}^* a_{\beta^*} \neq \mathbf{0}_p$. Thus $a_{\beta^*} \neq 0$ and $\tilde{\boldsymbol{\beta}}^* \neq \mathbf{0}_p$. □

Remark 3.13. *If all of the assumptions imposed in the Theorem 3.12 are satisfied and* $\operatorname{Cov}(\boldsymbol{B}^T \boldsymbol{X}, Y) \neq \boldsymbol{0}_k$ *then we have* $a_{\beta^*} > 0$. *We observe that* $\tilde{\boldsymbol{\beta}}^* \neq \boldsymbol{0}_p$ *and in view of Theorem 3.12 we get:*

$$\boldsymbol{\eta} = \frac{a_{\boldsymbol{B}}}{a_{\beta^*}} \neq \boldsymbol{0}_k.$$

From (3.20) and from Lemma A.43 we obtain

$$0 \leq \operatorname{Cov}(\tilde{\boldsymbol{\beta}}^{*T} \tilde{\boldsymbol{X}}, q_L(\beta_0^* + \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{\beta}}^*)) = \tilde{\boldsymbol{\beta}}^{*T} \mathbf{\Sigma} \tilde{\boldsymbol{\beta}}^* a_{\beta^*} = ||\mathbf{\Sigma}^{\frac{1}{2}} \tilde{\boldsymbol{\beta}}^*||_2^2 a_{\beta^*}.$$

Thus $a_{\beta^*} > 0$ *as* $a_{\beta^*} \neq 0$ *and we obtain* $s^* \neq \emptyset$.

Remark 3.14. *Theorem 3.12 holds also for the model without intercept. We have in this case:*

$$a_{\boldsymbol{B}} = (\operatorname{Var}(\tilde{\boldsymbol{B}}^T \tilde{\boldsymbol{X}}))^{-1} \mathbb{E} \tilde{\boldsymbol{B}}^T \tilde{\boldsymbol{X}} Y$$

and

$$a_{\beta^*} = (\operatorname{Var}(\tilde{\boldsymbol{\beta}}^{*T} \tilde{\boldsymbol{X}}))^{-1} \mathbb{E} \tilde{\boldsymbol{\beta}}^{*T} \tilde{\boldsymbol{X}} Y.$$

Lemma 3.15. *If* q *is differentiable,* $\tilde{\boldsymbol{X}}$ *follow normal distribution with* $\mathbf{\Sigma} > 0$, $\operatorname{rank} \tilde{\boldsymbol{B}} = k$, $\mathbb{E}||Dq(\boldsymbol{B}^T \boldsymbol{X})||_2 < \infty$, *then* $a_{\boldsymbol{B}} = \mathbb{E} Dq(\boldsymbol{B}^T \boldsymbol{X})$ *and* $a_{\beta^*} = \mathbb{E} q_L'(\boldsymbol{\beta}^{*T} \boldsymbol{X}) \in (0, 1/4)$.

Proof. Proof is similar to the proof of Lemma 2.23 (we replace $\boldsymbol{\beta}$ by $\mathbf{B}$ and use Lemma A.46 instead of Lemma A.45). □

Remark 3.16. *If all of assumptions of Lemma 3.15 are satisfied then*

$$\boldsymbol{\eta} = \frac{\mathbb{E} Dq(\boldsymbol{B}^T \boldsymbol{X})}{\mathbb{E} q_L'(\boldsymbol{\beta}^{*T} \boldsymbol{X})} \tag{3.21}$$

and each coordinate of $\boldsymbol{\eta}$ *satisfies:*

$$|\eta_i| > 4 |\mathbb{E} D_i q(\boldsymbol{B}^T \boldsymbol{X})|. \tag{3.22}$$

Proposition 3.17. *Let* $\boldsymbol{X} = (X_0, \tilde{\boldsymbol{X}}_1^T, \tilde{\boldsymbol{X}}_2^T)^T$, $\boldsymbol{B} = (\boldsymbol{B}_0, \tilde{\boldsymbol{B}}_1^T, \tilde{\boldsymbol{B}}_2^T)^T$ *and* $\tilde{\boldsymbol{B}}_1 \in \mathbb{R}^{m \times k}$, $\tilde{\boldsymbol{X}}_1 \in \mathbb{R}^m$, $\tilde{\boldsymbol{B}}_2 \in \mathbb{R}^{(p-m) \times k}$, $\tilde{\boldsymbol{X}}_2 \in \mathbb{R}^{p-m}$, $\operatorname{Cov}(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{X}}_j) = \mathbf{\Sigma}_{ij}$ *for* $i, j = 1, 2$. *Suppose that logistic model* $Y \sim \beta_0^* + \tilde{\boldsymbol{X}}_1^T \tilde{\boldsymbol{\beta}}_1^*$ *with omitted* $\tilde{\boldsymbol{X}}_2$ *variables is fitted and* $(\beta_0^*, \boldsymbol{\beta}_1^{*T})^T$ *is the corresponding projection. If* $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$, $\mathbf{\Sigma}_{11} > 0$, *LRC(*$\boldsymbol{B}$*) and LRC(*$(\beta_0^*, \tilde{\boldsymbol{\beta}}_1^{*T})^T$*) hold and* $\operatorname{Cov}(\boldsymbol{B}^T \boldsymbol{X}, Y) \neq \boldsymbol{0}_k$, *then we have*

$$\tilde{\boldsymbol{\beta}}_1^* = (\tilde{\boldsymbol{B}}_1 + \mathbf{\Sigma}_{11}^{-1} \mathbf{\Sigma}_{12} \tilde{\boldsymbol{B}}_2) \boldsymbol{\eta}, \tag{3.23}$$

where $\boldsymbol{\eta} = \frac{a_{\boldsymbol{B}}}{a_{\beta_1^*}} \neq \boldsymbol{0}_k$ *and*

$$a_{\beta_1^*} = \frac{\operatorname{Cov}(Y, \tilde{\boldsymbol{X}}_1^T \tilde{\boldsymbol{\beta}}_1^*)}{\operatorname{Var}(\tilde{\boldsymbol{X}}_1^T \tilde{\boldsymbol{\beta}}_1^*)} = \frac{\operatorname{Cov}(q_L(\beta_0^* + \tilde{\boldsymbol{X}}_1^T \tilde{\boldsymbol{\beta}}_1^*), \tilde{\boldsymbol{X}}_1^T \tilde{\boldsymbol{\beta}}_1^*)}{\operatorname{Var}(\tilde{\boldsymbol{X}}_1^T \tilde{\boldsymbol{\beta}}_1^*)}. \tag{3.24}$$

Proof. Analogously as in Theorem 3.12, we obtain the equations:

$$\mathrm{Cov}(\tilde{\mathbf{X}}_1, q(\mathbf{B}_0 + \tilde{\mathbf{B}}_1^T\tilde{\mathbf{X}}_1 + \tilde{\mathbf{B}}_2^T\tilde{\mathbf{X}}_2)) = \mathrm{Cov}(\tilde{\mathbf{X}}_1, \tilde{\mathbf{B}}_1^T\tilde{\mathbf{X}}_1 + \tilde{\mathbf{B}}_2^T\tilde{\mathbf{X}}_2)a_{\mathbf{B}},$$
$$\mathrm{Cov}(\tilde{\mathbf{X}}_1, q_L(\beta_0^* + \tilde{\boldsymbol{\beta}}_1^{*T}\tilde{\mathbf{X}}_1)) = \mathrm{Cov}(\tilde{\mathbf{X}}_1, \tilde{\boldsymbol{\beta}}_1^{*T}\tilde{\mathbf{X}}_1)a_{\boldsymbol{\beta}_1^*}.$$

Hence from normal equations 3.10 we have

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

what can be simplified to

$$\boldsymbol{\Sigma}_{11}\tilde{\boldsymbol{\beta}}_1^* a_{\boldsymbol{\beta}_1^*} = (\boldsymbol{\Sigma}_{11}\tilde{\mathbf{B}}_1 + \boldsymbol{\Sigma}_{12}\tilde{\mathbf{B}}_2)a_{\mathbf{B}}. \tag{3.25}$$

As $\boldsymbol{\Sigma}_{11}$ is invertible, we conclude that $a_{\boldsymbol{\beta}_1^*}$ is non-zero similarly as in Theorem 3.12. Multiplying both sides of Equation (3.25) by $(a_{\boldsymbol{\beta}_1^*}\boldsymbol{\Sigma}_{11})^{-1}$, we obtain the conclusion. □

Remark 3.18. *If in Proposition 3.17 we assume additionally that $\boldsymbol{\Sigma}_{12} = \boldsymbol{O}_{m\times(p-m)}$, then $\tilde{\boldsymbol{\beta}}_1^* = \tilde{\boldsymbol{B}}_1\boldsymbol{\eta}$. For independent $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$ we thus obtain a complementary conclusion to that of Theorem 3.1.*

3.3. $\boldsymbol{\beta}^*$ as first canonical vector

Lemma 3.19. *If $\boldsymbol{X}$ is a random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2 < \infty$, $\mathbb{E}\tilde{\boldsymbol{X}} = \boldsymbol{\mu}$ and LRC($\boldsymbol{B}$) is satisfied, then for $\boldsymbol{Z} = \tilde{\boldsymbol{X}} - \boldsymbol{\mu}$ we have:*

$$\mathbb{E}(\boldsymbol{Z}|\tilde{\boldsymbol{B}}^T\boldsymbol{Z}) = \boldsymbol{H}\tilde{\boldsymbol{B}}^T\boldsymbol{Z}.$$

Proof. Using Lemma 3.10, we have:

$$\begin{aligned}\mathbb{E}(\mathbf{Z}|\tilde{\mathbf{B}}^T\mathbf{Z} = \mathbf{d}) &= \mathbb{E}(\tilde{\mathbf{X}} - \boldsymbol{\mu}|\tilde{\mathbf{B}}^T\tilde{\mathbf{X}} = \tilde{\mathbf{B}}^T\boldsymbol{\mu} + \mathbf{d}) = \mathbf{h}_0 + \mathbf{H}(\tilde{\mathbf{B}}^T\boldsymbol{\mu} + \mathbf{d}) - \boldsymbol{\mu} \\ &= \mathbf{h}_0 - \boldsymbol{\mu} + \mathbf{H}\tilde{\mathbf{B}}^T\boldsymbol{\mu} + \mathbf{H}\mathbf{d} = (\mathbf{I}_p - \mathbf{H}\tilde{\mathbf{B}}^T)\boldsymbol{\mu} - \boldsymbol{\mu} + \mathbf{H}\tilde{\mathbf{B}}^T\boldsymbol{\mu} + \mathbf{H}\mathbf{d} = \mathbf{H}\mathbf{d}.\end{aligned}$$

□

It turns out that when LRC($\mathbf{B}$) and LRC($\boldsymbol{\beta}^*$) are satisfied, then the direction of the first canonical vector defined in (3.26) is the same as direction of $\tilde{\boldsymbol{\beta}}^*$ in the case of logistic loss, what follows from the theorem below. This sheds a new light on known effectiveness of canonical analysis in classification problems.

Theorem 3.20. *Let $\boldsymbol{X}$ be a random vector such that $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$, $\operatorname{rank}\tilde{\boldsymbol{B}} = k$, LRC($\boldsymbol{B}$) is satisfied and*

$$\boldsymbol{w} = \underset{\boldsymbol{v}\in\mathbb{R}^p\setminus\{0\}}{\arg\max}\,\frac{\boldsymbol{v}^T\boldsymbol{\Gamma}\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}, \tag{3.26}$$

where $\boldsymbol{\Gamma} = \mathrm{Var}(\mathbb{E}(\tilde{\boldsymbol{X}}|Y))$. If $\mathrm{Cov}(\tilde{\boldsymbol{X}}, Y) \neq \boldsymbol{0}_p$ and $q(\boldsymbol{B}^T\boldsymbol{X}) \in (0,1)$ $\mathbb{P}_{\boldsymbol{X}}$ a.e., then $\boldsymbol{w} = d\cdot\tilde{\boldsymbol{B}}a_{\boldsymbol{B}}$ for some $d \neq 0$, where $a_{\boldsymbol{B}}$ is defined in (3.17).

Proof. Let $\mathbf{Z} = \tilde{\mathbf{X}} - \mathbb{E}\tilde{\mathbf{X}}$ and $\boldsymbol{\mu}_i = \mathbb{E}(\mathbf{Z}|Y = i)$ for $i = 0, 1$.
Then it is easy to check:

$$\boldsymbol{\mu}_1 = \frac{\mathbb{E}\mathbf{Z}q(\mathbf{B}^T\mathbf{X})}{\mathbb{E}q(\mathbf{B}^T\mathbf{X})}, \quad \boldsymbol{\mu}_0 = \frac{-\mathbb{E}\mathbf{Z}q(\mathbf{B}^T\mathbf{X})}{1 - \mathbb{E}q(\mathbf{B}^T\mathbf{X})}.$$

As Y is binary, we have:

$$\operatorname{Var} Y = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = \mathbb{E}Y - (\mathbb{E}Y)^2 = \mathbb{E}q(\mathbf{B}^T\mathbf{X})(1 - \mathbb{E}q(\mathbf{B}^T\mathbf{X})).$$

Moreover we obtain:

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 = \mathbb{E}\mathbf{Z}q(\mathbf{B}^T\mathbf{X}) \left(\frac{1 - \mathbb{E}q(\mathbf{B}^T\mathbf{X}) + \mathbb{E}q(\mathbf{B}^T\mathbf{X})}{\mathbb{E}q(\mathbf{B}^T\mathbf{X})(1 - \mathbb{E}q(\mathbf{B}^T\mathbf{X}))} \right) = \frac{\mathbb{E}\mathbf{Z}q(\mathbf{B}^T\mathbf{X})}{\mathbb{E}q(\mathbf{B}^T\mathbf{X})(1 - \mathbb{E}q(\mathbf{B}^T\mathbf{X}))}.$$

Hence:

$$\begin{aligned} \boldsymbol{\Gamma} = \operatorname{Var}(\mathbb{E}(\mathbf{Z}|Y)) = \operatorname{Var}(\boldsymbol{\mu}_0 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)Y) &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \operatorname{Var} Y (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \\ &= \frac{\mathbb{E}\mathbf{Z}q(\mathbf{B}^T\mathbf{X}) \cdot (\mathbb{E}\mathbf{Z}q(\mathbf{B}^T\mathbf{X}))^T}{\mathbb{E}q(\mathbf{B}^T\mathbf{X})(1 - \mathbb{E}q(\mathbf{B}^T\mathbf{X}))}. \end{aligned}$$

Now observe that from LRC($\mathbf{B}$), Theorem 3.10 and Lemma 3.19 it follows that:

$$\begin{aligned} \mathbb{E}\mathbf{Z}q(\mathbf{B}^T\mathbf{X}) &= \mathbb{E}(\mathbb{E}(\mathbf{Z}q(\mathbf{B}^T\mathbf{X})|\tilde{\mathbf{B}}^T\mathbf{Z})) = \mathbf{H}\mathbb{E}\tilde{\mathbf{B}}^T\mathbf{Z}q(\mathbf{B}^T\mathbf{X}) \\ &= \boldsymbol{\Sigma}\tilde{\mathbf{B}}(\tilde{\mathbf{B}}^T\boldsymbol{\Sigma}\tilde{\mathbf{B}})^{-1}\mathbb{E}(\tilde{\mathbf{B}}^T\tilde{\mathbf{X}} - \mathbb{E}\tilde{\mathbf{B}}^T\tilde{\mathbf{X}})q(\mathbf{B}^T\mathbf{X}) = \boldsymbol{\Sigma}\tilde{\mathbf{B}}a_{\mathbf{B}}. \end{aligned}$$

Thus we have:

$$\boldsymbol{\Gamma} = \frac{1}{\mathbb{E}q(\mathbf{B}^T\mathbf{X})(1 - \mathbb{E}q(\mathbf{B}^T\mathbf{X}))}\boldsymbol{\Sigma}\tilde{\mathbf{B}}a_{\mathbf{B}}a_{\mathbf{B}}^T\tilde{\mathbf{B}}^T\boldsymbol{\Sigma},$$

therefore

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma} = \frac{1}{\mathbb{E}q(\mathbf{B}^T\mathbf{X})(1 - \mathbb{E}q(\mathbf{B}^T\mathbf{X}))}\tilde{\mathbf{B}}a_{\mathbf{B}}a_{\mathbf{B}}^T\tilde{\mathbf{B}}^T\boldsymbol{\Sigma}.$$

Let $\mathbf{a} = \tilde{\mathbf{B}}a_{\mathbf{B}}$, $\mathbf{b} = \boldsymbol{\Sigma}\mathbf{a}$. Clearly $\mathbf{a}, \mathbf{b}$ are vectors and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}$ has the form :

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma} = c\mathbf{a}\mathbf{b}^T,$$

where $c = (\mathbb{E}q(\mathbf{B}^T\mathbf{X})(1 - \mathbb{E}q(\mathbf{B}^T\mathbf{X})))^{-1}$. Hence we obtain

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\mathbf{a} = c\mathbf{a}^T\mathbf{b}\mathbf{a}$$

and $c\mathbf{a}^T\mathbf{b} = ca_{\mathbf{B}}^T\tilde{\mathbf{B}}\boldsymbol{\Sigma}\tilde{\mathbf{B}}a_{\mathbf{B}} > 0$ ($\tilde{\mathbf{B}}\boldsymbol{\Sigma}\tilde{\mathbf{B}}$ is positive definite, as $\operatorname{rank}\tilde{B} = k$ and $\boldsymbol{\Sigma} > 0$). This means that $\mathbf{a}$ is the eigenvector of the matrix $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}$ corresponding to the largest eigenvalue as matrix $\mathbf{a}\mathbf{b}^T$ has rank 1 and remaining eigenvalues are equal to 0. Hence $\mathbf{w} = d\mathbf{a}$ for $d \neq 0$. □

3.4. Logistic loss - additive binary model

In this section we will consider properties of projections for a special case of (3.1), namely

$$q(\mathbf{B}^T\mathbf{X}) = \lambda_1 q_1(\boldsymbol{\beta}_1^T\mathbf{X}) + \cdots + \lambda_k q_k(\boldsymbol{\beta}_k^T\mathbf{X}), \tag{3.27}$$

where $\mathbf{X} \in \mathbb{R}^p$, $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k] \in \mathbb{R}^{p\times k}$, $\lambda_i \geq 0$, $i = 1, \ldots, k$, $\sum_{i=1}^k \lambda_i = 1$ and $q_i \colon \mathbb{R} \to (0, 1)$ are differentiable. Moreover, we will assume that q_i are strictly increasing. Such model will be called an additive binary model. We assume that $\boldsymbol{\beta}_i$ do not contain intercept, as in the opposite case we may define $\tilde{q}_i$ as $\tilde{q}_i(x) = q_i(\beta_{i0} + x)$ instead of q_i. Note that a simple example of such model is a logistic mixture, where $q_i(s) = q_L(s)$ for all $i = 1, \ldots, n$.

Model (3.27) has the following natural interpretation. Namely, assume that in addition to $\mathbf{X}$ a discrete random variable I is observed which is independent of $\mathbf{X}$, $P(I = i) = \lambda_i$, $i = 1, \ldots, k$ and given $\mathbf{X} = \mathbf{x}$ and $I = i$ we have

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}, I = i) = q_i(\boldsymbol{\beta}_i^T \mathbf{x}), \tag{3.28}$$

for $i = 1, \ldots, k$. Thus averaging over I we obtain that $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$ is given by (3.27).

We will consider normal predictors $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} > 0$. Note that in this case LRC is satisfied for every $\mathbf{B} \in \mathbb{R}^{p\times k}$ with $\operatorname{rank} \mathbf{B} = k$ and $k < p$. We consider the case when logistic model without intercept $\mathbb{P}(Y = 1|\mathbf{X}) = q_L(\boldsymbol{\beta}^{*T}\mathbf{X})$ is fitted. This corresponds to the situation when q in (3.1) is calibrated in such a way that $q(0, \ldots, 0) = 0.5$. We note that assertions of the previous sections in this chapter hold also in this case and vector $\boldsymbol{\beta}^*$ is a linear combination of $\boldsymbol{\beta}_i$s.

When the logistic model with the intercept is fitted, Theorem 3.21 will hold and validity of the remaining results is still an open question. The proof of Lemma 3.22 uses the implicit function theorem and the derivation there relies crucially on the lack of intercept in $\boldsymbol{\beta}^*$.

For the additive binary model (3.27) we prove that the coefficients η_i of the combination in (3.5) are non-negative and establish upper bounds for them (cf. Theorem 3.23). In particular, when predictors are normal and $\boldsymbol{\beta}_i^T\mathbf{X}$ have the same variances, in the case of logistic mixture the bounds imply that $\eta_i \leq \lambda_i$. Moreover, we prove that the variance of $\boldsymbol{\beta}^{*T}\mathbf{X}$ is not larger than maximal variance of the projected linear combinations for the corresponding univariate problems.

Let

$$\mathbf{U} = (U_1, \ldots, U_k)^T = (\boldsymbol{\beta}_1^T\mathbf{X}, \ldots, \boldsymbol{\beta}_k^T\mathbf{X})^T.$$

We define D_i as the unique solutions of equations:

$$D_i = \frac{\mathbb{E}q_i'(U_i)}{\mathbb{E}q_L'(D_iU_i)}, \tag{3.29}$$

where $i = 1, \ldots, k$ and q_L is logistic function. Observe that existence and uniqueness of D_i follows from the following reasoning. Consider binary model $\mathbb{P}(Y = 1|\mathbf{X}) = q_i(\boldsymbol{\beta}_i^T\mathbf{X}) = q_i(U_i)$ and its projection on logistic model with pertaining vector $\boldsymbol{\beta}_i^*$. Then unique $\boldsymbol{\beta}_i^*$ exists in view of Remark A.12, as $\boldsymbol{\Sigma} > 0$ and $q_i(x) \in (0, 1)$ for $x \in \mathbb{R}$. This means, that in view of Lemma 2.23 and (2.27) $\boldsymbol{\beta}_i^* = \eta\boldsymbol{\beta}_i$ and:

$$\eta = \frac{\mathbb{E}q_i'(U_i)}{\mathbb{E}q_L'(\eta U_i)}. \tag{3.30}$$

Hence $\eta = D_i$ from uniqueness of η. Thus vectors $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_i^*$ are proportional and D_i are the constants of proportionality in the univariate projection problem. We first prove

Theorem 3.21. *Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} > 0$, $\operatorname{rank} \boldsymbol{B} = k$ and Y given $\boldsymbol{X}$ follows the conditional distribution defined in (3.27), where q_i are strictly increasing and differentiable such that $\mathbb{E}||q_i(\boldsymbol{\beta}_i^T \boldsymbol{X})\boldsymbol{X}||_2 < \infty$, $\mathbb{E}q_i'(\boldsymbol{\beta}_i^T \boldsymbol{X}) < \infty$. If $\boldsymbol{\beta}^* = \eta_1 \boldsymbol{\beta}_1 + \ldots + \eta_k \boldsymbol{\beta}_k$ then*

$$\eta_i = \lambda_i \frac{\mathbb{E}q_i'(U_i)}{\mathbb{E}q_L'(\boldsymbol{\eta}^T \boldsymbol{U})}, \tag{3.31}$$

and $\eta_i \geq 0$.

Proof. Normal equations (3.9) for $\eta_1, \ldots, \eta_k$ in view of $\boldsymbol{\beta}^{*T}\mathbf{X} = \sum_{i=1}^k \eta_i \boldsymbol{\beta}_i^T \mathbf{X}$ have the form:

$$\mathbb{E}q_L(\eta_1 \boldsymbol{\beta}_1^T \mathbf{X} + \ldots + \eta_k \boldsymbol{\beta}_k^T \mathbf{X})\mathbf{X} = \sum_{i=1}^k \lambda_i \mathbb{E}q_i(\boldsymbol{\beta}_i^T \mathbf{X})\mathbf{X}.$$

After multiplying these equations by matrix $\mathbf{B}^T$ and using the definition of $\mathbf{U}$, we obtain:

$$\mathbb{E}q_L(\boldsymbol{\eta}^T \mathbf{U})\mathbf{U} = \sum_{i=1}^k \lambda_i \mathbb{E}q_i(U_i)\mathbf{U}. \tag{3.32}$$

It follows from Stein's Lemma A.46 applied componentwise that (3.32) is equivalent to

$$\boldsymbol{\Sigma}_{\mathbf{U}} \boldsymbol{\eta} \mathbb{E}q_L'(\boldsymbol{\eta}^T \mathbf{U}) = \boldsymbol{\Sigma}_{\mathbf{U}} \mathbf{w},$$

where $\boldsymbol{\Sigma}_{\mathbf{U}} = \operatorname{Var} \mathbf{U} = \mathbf{B}^T \boldsymbol{\Sigma} \mathbf{B}$ and $\mathbf{w} = (\lambda_1 \mathbb{E}q_1'(U_1), \ldots, \lambda_k \mathbb{E}q_k'(U_k))^T$. Since $\boldsymbol{\Sigma}_{\mathbf{U}} > 0$ equality (3.31) follows.

From the equation above we observe that $\eta_i \geq 0$ and $\eta_i = 0$ only when $\lambda_i = 0$. Hence when $\lambda_i = 1$ for some i, then $\eta_j = 0$ for $j \neq i$ and from uniqueness of $\boldsymbol{\eta}$ we have $\eta_i = D_i$, where D_i is the constant defined in the equation (3.29). □

We state now the crucial lemma from which upper bounds on the coefficients η_i follow.

Lemma 3.22. *Assume that conditions of Theorem 3.21 are satisfied. Then*

$$\sum_{i=1}^k \frac{\eta_i}{\mathbb{E}q_i'(U_i)} \leq \max_{i=1,\ldots,k} \left(\frac{D_i}{\mathbb{E}q_i'(U_i)} \right) = \max_{i=1,\ldots,k} \left(\frac{1}{\mathbb{E}q_L'(D_i U_i)} \right). \tag{3.33}$$

Proof. Since it follows from (3.31) that $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T$ exists and is unique for each $\lambda_1, \ldots, \lambda_k$, we can consider $\boldsymbol{\eta}$ as a function of $\lambda_1, \ldots, \lambda_{k-1}$ as $\lambda_k = 1 - \sum_{i=1}^{k-1} \lambda_i$. We will use the implicit function theorem to prove the lemma.

Let us observe that the theorem is true for $k_0 = 1$. Now assume that it holds for $k_0 = k - 1 \geq 1$ and we proceed by induction. Let

$$B_k = \{(\lambda_1, \ldots, \lambda_{k-1}) \in \mathbb{R}^{k-1} : \forall i : \ \lambda_i \geq 0, \sum_{i=1}^{k-1} \lambda_i \leq 1\}.$$

Consider the following function $F\colon \ B_k \times \mathbb{R}^k \to \mathbb{R}^k$, where

$$F(\lambda_1, \ldots, \lambda_{k-1}, \eta_1, \ldots, \eta_k) = \mathbb{E}q_L(\boldsymbol{\eta}^T \mathbf{U})\mathbf{U} - \sum_{i=1}^{k-1} \lambda_i \mathbb{E}q_i(U_i)\mathbf{U} - \mathbb{E}q_k(U_k)\mathbf{U} \left(1 - \sum_{i=1}^{k-1} \lambda_i\right).$$

Function $F = (F_1, \ldots, F_k)$ equals the difference of both sides of (3.9) after substituting $1 - \sum_{i=1}^{k-1} \lambda_i$ for λ_k. We have for $m = 1, \ldots, k$:

$$\frac{\partial F_m}{\partial \lambda_j} = -\mathbb{E}q_j(U_j)U_m + \mathbb{E}q_k(U_k)U_m, \quad \frac{\partial F_m}{\partial \eta_j} = \mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U})U_m U_j.$$

We rewrite the above equations in matrix form:

$$D_{\boldsymbol{\eta}}F = \mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U})\mathbf{U}\mathbf{U}^T, \quad D_{\lambda_j}F = \mathbb{E}q_k(U_k)\mathbf{U} - \mathbb{E}q_j(U_j)\mathbf{U}.$$

As $\mathbb{E}\mathbf{U}\mathbf{U}^T > 0$, we observe that $D_{\boldsymbol{\eta}}F > 0$. This means that we can use the implicit function theorem to obtain:

$$D_{\boldsymbol{\eta}}F \cdot D_{\lambda_j}\boldsymbol{\eta} = -D_{\lambda_j}F. \tag{3.34}$$

Let $\mathbf{V} = \boldsymbol{\Sigma}_{\mathbf{U}}^{-\frac{1}{2}}\mathbf{U}$, where $\boldsymbol{\Sigma}_{\mathbf{U}} = \mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B}$ is the covariance matrix of $\mathbf{U}$. The above substitution and Lemma A.50 gives:

$$\begin{aligned} D_{\boldsymbol{\eta}}F &= \mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U})\mathbf{U}\mathbf{U}^T = \boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\mathbb{E}q'_L((\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta})^T\mathbf{V})\mathbf{V}\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}} \\ &= \boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\left(\mathbb{E}q'_L((\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta})^T\mathbf{V})\mathbf{I}_k + \mathbb{E}q'''_L((\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta})^T\mathbf{V})(\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta})(\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta})^T\right)\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}. \end{aligned} \tag{3.35}$$

It follows from the structure of matrix $D_{\boldsymbol{\eta}}F$ in (3.35) that it has the following eigenvalues: $a_1 = \ldots = a_{k-1} = \mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U}), a_k = \mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U}) + \mathbb{E}q'''_L(\boldsymbol{\eta}^T\mathbf{U})||\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta}||^2$, which are positive by positive definiteness of matrix $D_{\boldsymbol{\eta}}F$. Moreover, using formula

$$(\mathbf{I} + \lambda\mathbf{x}\mathbf{x}^T)^{-1} = \mathbf{I} - \frac{\lambda}{1 + ||\mathbf{x}||^2\lambda}\mathbf{x}\mathbf{x}^T,$$

we have:

$$\begin{aligned} (D_{\boldsymbol{\eta}}F)^{-1} &= \frac{\boldsymbol{\Sigma}_{\mathbf{U}}^{-\frac{1}{2}}}{\mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U})} \cdot \left(\mathbf{I}_k - \frac{\mathbb{E}q'''_L(\boldsymbol{\eta}^T\mathbf{U})}{\mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U}) + ||\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta}||^2\mathbb{E}q'''_L(\boldsymbol{\eta}^T\mathbf{U})}(\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta})(\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta})^T\right)\boldsymbol{\Sigma}_{\mathbf{U}}^{-\frac{1}{2}} \\ &= \frac{1}{\mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U})}\left(\boldsymbol{\Sigma}_{\mathbf{U}}^{-1} - \frac{\mathbb{E}q'''_L(\boldsymbol{\eta}^T\mathbf{U})}{\mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U}) + ||\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta}||^2\mathbb{E}q'''_L(\boldsymbol{\eta}^T\mathbf{U})} \cdot \boldsymbol{\eta}\boldsymbol{\eta}^T\right). \end{aligned} \tag{3.36}$$

From the Stein's Lemma A.45 we obtain:

$$\begin{aligned} D_{\lambda_1}F &= -\mathbb{E}q_1(U_1)\mathbf{U} + \mathbb{E}q_k(U_k)\mathbf{U} = -\mathbb{E}q'_1(U_1)\operatorname{Cov}(\mathbf{U}, U_1) + \mathbb{E}q'_k(U_k)\operatorname{Cov}(\mathbf{U}, U_k) \\ &= -\mathbb{E}q'_1(U_1)\boldsymbol{\Sigma}_{\mathbf{U}}\mathbf{e}_1 + \mathbb{E}q'_k(U_k)\boldsymbol{\Sigma}_{\mathbf{U}}\mathbf{e}_k = \boldsymbol{\Sigma}_{\mathbf{U}}(-\mathbb{E}q'_1(U_1)\mathbf{e}_1 + \mathbb{E}q'_k(U_k)\mathbf{e}_k), \end{aligned}$$

where $\mathbf{e}_i$ is i^{th} vector of the standard basis in $\mathbb{R}^k$ for $i = 1, \ldots, k$. Hence:

$$\begin{aligned} D_{\lambda_1}\boldsymbol{\eta} &= -(D_{\boldsymbol{\eta}}F)^{-1} \cdot D_{\lambda_1}F \\ &= \frac{1}{\mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U})}\left(\mathbf{I}_k - \frac{\mathbb{E}q'''_L(\boldsymbol{\eta}^T\mathbf{U})}{\mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U}) + ||\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta}||^2\mathbb{E}q'''_L(\boldsymbol{\eta}^T\mathbf{U})}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\mathbf{U}}\right) \\ &\quad \cdot (\mathbb{E}q'_1(U_1)\mathbf{e}_1 - \mathbb{E}q'_k(U_k)\mathbf{e}_k). \end{aligned} \tag{3.37}$$

Let:

$$\mathbf{v} = \left(\frac{1}{\mathbb{E}q'_1(U_1)}, \frac{1}{\mathbb{E}q'_2(U_2)}, \ldots, \frac{1}{\mathbb{E}q'_k(U_k)}\right)^T \in \mathbb{R}^k.$$

Then we obtain:

$$\sum_{i=1}^{k} \frac{1}{\mathbb{E}q'_i(U_i)}\frac{\partial \eta_i}{\partial \lambda_1} = \mathbf{v}^T D_{\lambda_1}\boldsymbol{\eta} = \frac{1}{\mathbb{E}q'_L(\boldsymbol{\eta}^T\mathbf{U})} \cdot \left(\underbrace{\mathbf{v}^T(\mathbb{E}q'_1(U_1)\mathbf{e}_1 - \mathbb{E}q'_k(U_k)\mathbf{e}_k)}_{0}\right.$$

$$
\left.
-\frac{\mathbb{E}q_L'''(\boldsymbol{\eta}^T\mathbf{U})}{\mathbb{E}q_L'(\boldsymbol{\eta}^T\mathbf{U}) + ||\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta}||^2\mathbb{E}q_L'''(\boldsymbol{\eta}^T\mathbf{U})}(\mathbf{v}^T\boldsymbol{\eta})\cdot\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\mathbf{U}}(\mathbb{E}q_1'(U_1)\mathbf{e}_1 - \mathbb{E}q_k'(U_k)\mathbf{e}_k)\right)
$$

$$
= \underbrace{\frac{-\mathbb{E}q_L'''(\boldsymbol{\eta}^T\mathbf{U})}{\mathbb{E}q_L'(\boldsymbol{\eta}^T\mathbf{U})}}_{>0 \text{ (Lemma A.51)}} \cdot \underbrace{\frac{1}{\mathbb{E}q_L'(\boldsymbol{\eta}^T\mathbf{U}) + ||\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{\eta}||^2\mathbb{E}q_L'''(\boldsymbol{\eta}^T\mathbf{U})}}_{>0} \cdot \underbrace{\left(\sum_{i=1}^{k}\frac{\eta_i}{\mathbb{E}q_i'(U_i)}\right)}_{>0} \times
$$

$$
\times \left(\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\mathbf{U}}(\mathbb{E}q_1'(U_1)\mathbf{e}_1 - \mathbb{E}q_k'(U_k)\mathbf{e}_k)\right).
$$

From the above equation and in view of Lemma A.51 we observe that the sign of its left hand side is the same as the sign of the expression $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\mathbf{U}}(\mathbb{E}q_1'(U_1)\mathbf{e}_1 - \mathbb{E}q_k'(U_k)\mathbf{e}_k)$. Let $\boldsymbol{\Sigma}_{\mathbf{U}} = [\sigma_{ij}]_{i,j=1,\ldots,k}$. Then from equalities (3.31), $\lambda_k = 1 - \sum_{i=1}^{k-1}\lambda_i$ and from symmetry of $\boldsymbol{\Sigma}_{\mathbf{U}}$ we obtain the following:

$$
\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\mathbf{U}}(\mathbb{E}q_1'(U_1)\mathbf{e}_1 - \mathbb{E}q_k'(U_k)\mathbf{e}_k) = \frac{\lambda_1 C_1 + C_2}{\mathbb{E}q_L'(\boldsymbol{\eta}^T\mathbf{U})}, \tag{3.38}
$$

where

$$
C_1 = (\mathbb{E}q_1'(U_1))^2\sigma_{11} - 2\mathbb{E}q_1'(U_1)\mathbb{E}q_k'(U_k)\sigma_{1k} + (\mathbb{E}q_k'(U_k))^2\sigma_{kk}
$$

and

$$
C_2 = \sum_{i=2}^{k-1}\lambda_i\Big(\mathbb{E}q_i'(U_i)\mathbb{E}q_1'(U_1)\sigma_{i1} - \mathbb{E}q_i'(U_i)\mathbb{E}q_k'(U_k)\sigma_{ik} - \mathbb{E}q_1'(U_1)\mathbb{E}q_k'(U_k)\sigma_{1k}
$$

$$
+ (\mathbb{E}q_k'(U_k))^2\sigma_{kk}\Big) + \mathbb{E}q_1'(U_1)\mathbb{E}q_k'(U_k)\sigma_{1k} - (\mathbb{E}q_k'(U_k))^2\sigma_{kk}.
$$

From inequality $x^2 + y^2 \geq 2xy$ and positive definiteness of $\boldsymbol{\Sigma}_{\mathbf{U}}$ we have:

$$
(\mathbb{E}q_1'(U_1))^2\sigma_{11} + (\mathbb{E}q_k'(U_k))^2\sigma_{kk} \geq 2\mathbb{E}q_1'(U_1)\mathbb{E}q_k'(U_k)\sqrt{\sigma_{11}\sigma_{kk}} > 2\mathbb{E}q_1'(U_1)\mathbb{E}q_k'(U_k)\sigma_{1k}.
$$

Thus $C_1 > 0$. Hence we obtain that $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\mathbf{U}}(\mathbb{E}q_1'(U_1)\mathbf{e}_1 - \mathbb{E}q_k'(U_k)\mathbf{e}_k) > 0$ if and only if $\lambda_1 > h$, where $h = -C_2{C_1}^{-1}$. Therefore $\sum_{i=1}^{k}(\mathbb{E}q_i'(U_i))^{-1}\frac{\partial\eta_i}{\partial\lambda_1} > 0$ if and only if $\lambda_1 > h$. Thus function $\sum_{i=1}^{k}(\mathbb{E}q_i'(U_i))^{-1}\eta_i$ is increasing function of λ_1 for $\lambda_1 > h$ and decreasing for $\lambda_1 < h$. Because $(\lambda_1, \ldots, \lambda_{k-1}) \in B_k$, we have $\lambda_1 \in [0, 1 - \lambda_2 - \ldots - \lambda_{k-1}]$. This means that:

$$
\left(\sum_{i=1}^{k}\frac{\eta_i}{\mathbb{E}q_i'(U_i)}\right)(\lambda_1, \lambda_2, \ldots, \lambda_{k-1}) \leq \max\left\{\left(\sum_{i=1}^{k}\frac{\eta_i}{\mathbb{E}q_i'(U_i)}\right)(0, \lambda_2, \ldots, \lambda_{k-1}),\right.
$$

$$
\left.\left(\sum_{i=1}^{k}\frac{\eta_i}{\mathbb{E}q_i'(U_i)}\right)(1 - \lambda_2 - \ldots - \lambda_{k-1}, \lambda_2, \ldots, \lambda_{k-1})\right\}. \tag{3.39}
$$

But the right hand side of the above inequality by induction step is bounded from above by:

$$
\max\left\{\max_{i=2,\ldots,k}\left(\frac{D_i}{\mathbb{E}q_i'(U_i)}\right), \max_{i=1,\ldots,k-1}\left(\frac{D_i}{\mathbb{E}q_i'(U_i)}\right)\right\} = \max_{i=1,\ldots,k}\left(\frac{D_i}{\mathbb{E}q_i'(U_i)}\right)
$$

and we have finally proved inequality (3.33). □

Theorem 3.23. *Assume that conditions of Theorem 3.21 are satisfied,*

$$
D = \max_{i=1,\ldots,k}\left(\frac{D_i}{\mathbb{E}q_i'(U_i)}\right) = \max_{i=1,\ldots,k}\frac{1}{\mathbb{E}q_L'(D_iU_i)}
$$

and D_i are defined in (3.29). Then (3.33) is equivalent to:

$$\eta_i \leq \lambda_i D \mathbb{E} q_i'(U_i) \text{ for all } i \in \{1, \ldots, k\} \tag{3.40}$$

and to

$$\mathrm{Var}(\boldsymbol{\eta}^T \boldsymbol{U}) \leq \max_{i=1,\ldots,k} \mathrm{Var}(D_i U_i). \tag{3.41}$$

Proof. (3.33) implies (3.40), as from Lemma 3.22, (3.31) and $\sum_{i=1}^k \lambda_i = 1$ we have :

$$D = \max_{i=1,\ldots,k} \left(\frac{D_i}{\mathbb{E} q_i'(U_i)} \right) \geq \sum_{i=1}^k \frac{\eta_i}{\mathbb{E} q_i'(U_i)} = \sum_{i=1}^k \frac{\lambda_i}{\mathbb{E} q_L'(\boldsymbol{\eta}^T \mathbf{U})} = \frac{1}{\mathbb{E} q_L'(\boldsymbol{\eta}^T \mathbf{U})}.$$

Thus again from equality (3.31) and the above inequality we obtain:

$$\eta_i = \lambda_i \frac{\mathbb{E} q_i'(U_i)}{\mathbb{E} q_L'(\boldsymbol{\eta}^T \mathbf{U})} \leq \lambda_i \mathbb{E} q_i'(U_i) D.$$

Conversely, (3.40) implies (3.33) as

$$\sum_{i=1}^k \frac{\eta_i}{\mathbb{E} q_i'(U_i)} \leq \sum_{i=1}^k \lambda_i D = D.$$

Moreover, (3.33) is equivalent to (3.41) as in view of (3.29), (3.31) and (3.12) inequality (3.33) is equivalent to the following inequality:

$$\frac{1}{\mathbb{E} q_L'(\boldsymbol{\eta}^T \mathbf{U})} \leq \max_{i=1,\ldots,k} \frac{1}{\mathbb{E} q_L'(D_i U_i)}.$$

Function $h(\sigma) = \mathbb{E} q_L'(\sigma Z)$ is decreasing if $\sigma \geq 0$ and $Z \sim \mathcal{N}(0,1)$ in view of Lemma A.51, statement 4 as $h'(\sigma) = \mathbb{E} q_L''(\sigma Z) Z < 0$. Thus the last inequality implies (3.41). □

Observe that $\mathrm{Var}(\boldsymbol{\eta}^T \mathbf{U}) = \mathrm{Var}(\boldsymbol{\beta}^{*T} \mathbf{X})$ and $\mathrm{Var}(D_i U_i) = \mathrm{Var}(\boldsymbol{\beta}_i^{*T} \mathbf{X})$, where $\boldsymbol{\beta}_i^*$ are defined below (3.29). Thus (3.41) can be stated as

$$\mathrm{Var}(\boldsymbol{\beta}^{*T} \mathbf{X}) \leq \max_{i=1,\ldots,k} \mathrm{Var}(\boldsymbol{\beta}_i^{*T} \mathbf{X}).$$

Thus the variance of $\boldsymbol{\beta}^{*T} \mathbf{X}$ is not larger than the maximal variance of the projected linear combinations in the corresponding univariate problems. Another way of interpreting (3.40) is to say that contribution to $\boldsymbol{\beta}^*$ of i^{th} component $\boldsymbol{\beta}_i$ is bounded by the term proportional to $C_i \lambda_i$, where λ_i is the probability with which i^{th} model is chosen and $C_i = \mathbb{E} q_i'(U_i)$ depends only on the i^{th} model (cf. the discussion of the additive model below (3.27)). Also note that $\mathbb{E} q_L'(D_i U_i)$ is an averaged variance of the response in the univariate logistic model with parameter $\boldsymbol{\beta}_i^*$. Thus D^{-1} equals to the minimal averaged variability for such models.

Corollary 3.24. *If the assumptions of Lemma 3.22 hold and $q_1 = \ldots = q_k = q_L$, then for any $i \in \{1, \ldots, k\}$*

$$\eta_i \leq \lambda_i \frac{\mathbb{E} q_L'(U_i)}{\min\limits_{j=1,\ldots,k} \mathbb{E} q_L'(U_j)}.$$

Proof. Observe that D_i defined in (3.29) are equal to 1. This means that D in Theorem 3.23 satisfies:

$$D = \max_{j=1,\ldots,k} \left(\frac{1}{\mathbb{E} q_L'(U_j)} \right) = \frac{1}{\min\limits_{j=1,\ldots,k} \mathbb{E} q_L'(U_j)}$$

and the theorem follows. □

Definition 3.25. *We say that vector* $\boldsymbol{X} = (X_1, \ldots, X_m)^T$ *is balanced, if*

$$\operatorname{Var} X_1 = \ldots = \operatorname{Var} X_m = \sigma^2 < \infty.$$

Corollary 3.26. *If the assumptions of Lemma 3.22 hold,* $q_1 = \ldots = q_k$ *and the vector* $\boldsymbol{U}$ *is balanced, then for any* $i \in \{1, \ldots, k\}$

$$\eta_i \leq \lambda_i D_1.$$

Equality in these inequalities holds if and only if $\lambda_i = 1$ *for some* $i \in \{1, \ldots, k\}$:

Proof. From the fact that $\mathbf{U}$ is normally distributed and balanced it follows that $U_1 \stackrel{d}{=} \ldots \stackrel{d}{=} U_k$ and as $q_1 = \ldots = q_k$ we have $\mathbb{E}q_1'(U_1) = \ldots = \mathbb{E}q_k'(U_k)$. From the uniqueness of D_i, which satisfies the equation (3.31), we have $D_1 = \ldots = D_k$ and D in Theorem 3.23 satisfies:

$$D = \frac{D_1}{\mathbb{E}q_1'(U_1)}.$$

Hence from Theorem 3.23 we obtain $\eta_i \leq \lambda_i D_1$. The last statement of the theorem follows from the proof of Lemma 3.22, namely inequality (3.39) is strict when the inequality $0 < \lambda_1 < 1 - \sum_{j=2}^{k-1} \lambda_j$ holds. □

We call (3.27) a balanced additive logistic model when $q_1 = \ldots = q_k = q_L$ and vector $\mathbf{U}$ is balanced.

Corollary 3.27. *If the assumptions of Lemma 3.22 hold and (3.27) is a balanced additive logistic model then*

$$\forall i \in \{1, \ldots, k\} : \ \eta_i \leq \lambda_i.$$

Equality in these inequalities holds if and only if for some $i \in \{1, \ldots, k\}$ $\lambda_i = 1$.

Proof. From Corollary 3.26 we obtain:

$$\forall i \in \{1, \ldots, k\} : \ \eta_i \leq D_1 \lambda_i,$$

where the equality holds if and only if if for some $i \in \{1, \ldots, k\}$ $\lambda_i = 1$. However, as in proof of Theorem 3.24, we obtain $D_1 = 1$. □

Finally, we give examples of the situations when using the proved results it is possible to bound the norm of the vector $\boldsymbol{\beta}^*$ or its coefficients. The bounds on $||\boldsymbol{\beta}^*||_2$ may be useful when calculating its maximum likelihood estimator. Then optimisation of the likelihood may be restricted to the ball $B(0, c)$, where c is specified below.

Corollary 3.28. *Let* $p = k$, $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{0}_p, \sigma^2 \boldsymbol{I})$, $\lambda_i \geq 0$ *for* $i = 1, \ldots, p$, $\sum_{i=1}^p \lambda_i = 1$, $q_1 = \ldots = q_p$, q_1 *is differentiable, increasing, such that* $q_1'(x) < c q_L'(cx)$ *for every* $x \in \mathbb{R}$, $\boldsymbol{B} = \boldsymbol{I}_p$ *thus*

$$q(\boldsymbol{\beta}_1^T \boldsymbol{X}, \ldots, \boldsymbol{\beta}_p^T \boldsymbol{X}) = \lambda_1 q_1(X_1) + \ldots + \lambda_p q_1(X_p)$$

and assumptions of Lemma 3.22 are fulfilled. Then $D_i < c$ *and* $||\boldsymbol{\beta}^*||_2 < c$.

Proof. Suppose that $D_i \geq c$. Then we would have from equation (3.29) and Lemma A.51, statement 6:

$$\mathbb{E}q_1'(X_i) = D_i \mathbb{E}q_L'(D_i X_i) \geq c\mathbb{E}q_L'(cX_i) > \mathbb{E}q_1'(X_i).$$

Thus $D_i < c$ and from the Theorem 3.23 we obtain:

$$\sigma^2 ||\boldsymbol{\beta}^*||_2^2 = \mathrm{Var}(\boldsymbol{\beta}^{*T}\mathbf{X}) \leq \max_{i=1,\ldots,k} \mathrm{Var}(D_i X_i) = D_1^2 \sigma^2 < c^2 \sigma^2.$$

□

Corollary 3.29. *Under assumptions of the Corollary 3.27 we have for* $i = 1, \ldots, p$*:*

$$|\beta_i^*| \leq \max_{j=1,\ldots,k} |\beta_{ji}|. \tag{3.42}$$

Proof.

$$|\beta_i^*| = \left| \sum_{j=1}^{k} \eta_j \beta_{ji} \right| \leq \sum_{j=1}^{k} |\eta_j||\beta_{ji}| \leq \sum_{j=1}^{k} \lambda_j |\beta_{ji}| \leq \max_{j=1,\ldots,k} |\beta_{ji}|.$$

□

Inequality (3.42) means that in this case coefficients of $\boldsymbol{\beta}^*$ are shrunk in the sense specified above (and for $k = 1$ in the usual sense). Similar property for binary predictor was established in Gail et al. (1988).

Chapter 4

Properties of Lasso estimator in misspecified binary model

In this chapter we consider properties of Lasso estimator for a misspecified binary model. Study of properties of inferential procedures under misspecification goes back to White (1982) who considered consistency and asymptotic normality of ML estimators in such a case (see also Vuong (1989)). The subject resurfaced recently in the setting of high-dimensional regression models. Bühlmann et al. (2015) studied properties of debiased Lasso for misspecified linear model, see also Lu et al. (2012). Properties of Lasso estimator, in particular important separation property will be used in Chapter 5 to prove consistency of two-step selection procedure. We stress that some of the properties considered here, in particular separation property are known for deterministic predictors (see eg. Fan et al. (2014a)). Their modified versions proved here for random regressors required substantially different approaches and proofs.

We consider an i.i.d. random sample $(\mathbf{X}_1^{(n)}, Y_1^{(n)}), \ldots, (\mathbf{X}_n^{(n)}, Y_n^{(n)}) \stackrel{d}{=} (\mathbf{X}^{(n)}, Y^{(n)}) \in \mathbb{R}^{p_n+1} \times \{0,1\}$, where $\mathbf{X}^{(n)} \sim \mathbb{P}_{\mathbf{X}}$ and we analyse general binary model:

$$\mathbb{P}(Y^{(n)} = 1 | \mathbf{X}^{(n)} = \mathbf{x}) = q^{(n)}(\mathbf{x}), \tag{4.1}$$

where $\mathbf{x} \in \mathbb{R}^{p_n+1}$. We adopt triangular scenario: $\mathbf{X}_i^{(n)} = (X_{i0}^{(n)}, X_{i1}^{(n)}, \ldots, X_{ip_n}^{(n)})^T$, $X_{i0}^{(n)} \equiv 1$. Frequently considered scenario is the sequential one. In this case, when sample size n increases we observe values of new predictors additionally to the ones observed earlier. This is a special case of the above scheme as then $\mathbf{X}_i^{(n+1)} = (\mathbf{X}_i^{(n)T}, X_{i,p_n+1}, \ldots, X_{i,p_{n+1}})^T$. To simplify the notation, we will further write $q^{(n)} = q$, $\mathbf{X}_i^{(n)} = \mathbf{X}_i = (1, \tilde{\mathbf{X}}_i^T)^T, Y_i^{(n)} = Y_i$. We assume that coordinates X_{ij} of $\mathbf{X}_i$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p_n$ are subgaussian $Subg(\sigma_{jn}^2)$ with subgaussianity parameter $\sigma_{jn} > 0$ i.e. it holds that $\mathbb{E}e^{tX_{ij}} \leq e^{\frac{t^2\sigma_{jn}^2}{2}}$ for all $t \in \mathbb{R}$. For future reference let $s_n^2 = \max_{j=1,\ldots,p_n} \sigma_{jn}^2$ and we assume throughout that

$$\limsup_n s_n^2 < \infty.$$

In the sequential scenario this is equivalent to the assumption that all subgaussianity parameters are bounded from above. We assume existence and uniqueness of $\boldsymbol{\beta}^*$ which was defined in (1.8). Case of models with intercept (only for logistic loss) and without intercept are treated separately, as the assumptions needed differ significantly. We note that for distributions of $\tilde{\mathbf{X}}$ satisfying LRC, fitted model without intercept and model with intercept both yield the same active set of predictors for logistic and quadratic loss (see Remarks 2.22, 2.30).

Now let $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ denote matrix of experiment of dimension $n \times (p_n + 1)$ and let $\tilde{\mathbb{X}} = (\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_n)^T$. In this chapter we assume throughout that loss function is of the form:

$$l(\mathbf{b}, \mathbf{x}, y) = \rho(\mathbf{b}^T \mathbf{x}, y), \tag{4.2}$$

where $\rho : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ is some function, $\mathbf{b}, \mathbf{x} \in \mathbb{R}^{p_n+1}$, $y \in \{0, 1\}$. Further, we define empirical risk as:

$$R_n(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \rho(\mathbf{b}^T \mathbf{X}_i, Y_i). \tag{4.3}$$

We will denote by:

$$P_n(\mathbf{b}) = R_n(\mathbf{b}) + \lambda ||\tilde{\mathbf{b}}||_1 \tag{4.4}$$

l_1 penalized empirical risk, where $\mathbf{b} = (b_0, \tilde{\mathbf{b}}^T)^T$. In this chapter we will be interested in properties of minimizer $\hat{\boldsymbol{\beta}}_L$ of P_n in the fitted model with intercept:

$$\hat{\boldsymbol{\beta}}_L = \operatorname*{arg\,min}_{\mathbf{b} \in \mathbb{R}^{p_n+1}} P_n(\mathbf{b}) \tag{4.5}$$

and in the model without intercept:

$$\hat{\boldsymbol{\beta}}_{L,-0} = \operatorname*{arg\,min}_{\tilde{\mathbf{b}} \in \mathbb{R}^{p_n}} P_n((0, \tilde{\mathbf{b}}^T)^T). \tag{4.6}$$

In all of the theorems we assume that $\rho(\cdot, y)$ is convex function for all y and is bounded from below by $m \in \mathbb{R}$. These two properties assure that $\hat{\boldsymbol{\beta}}_L$ exists in view of Lemma A.3. If $\hat{\boldsymbol{\beta}}_L$ exists and $\rho(\cdot, y)$ is convex and differentiable function for all y, then $\hat{\boldsymbol{\beta}}_L$ satisfies Karush–Kuhn–Tucker conditions (in subgradient form):

$$0 \in \partial_{b_j} P_n(\hat{\boldsymbol{\beta}}_L) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \rho}{\partial b}(\hat{\boldsymbol{\beta}}_L^T \mathbf{X}_i, Y_i) X_{ij} + \lambda s_j(\hat{\boldsymbol{\beta}}_L) \right\} \text{ for } j \in \{1, \ldots, p_n\}, \tag{4.7}$$

and for $j = 0$:

$$0 \in \partial_{b_0} P_n(\hat{\boldsymbol{\beta}}_L) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \rho}{\partial b}(\hat{\boldsymbol{\beta}}_L^T \mathbf{X}_i, Y_i) \right\}, \tag{4.8}$$

where $\hat{\tilde{\boldsymbol{\beta}}}_L = (\hat{\beta}_{L,1}, \ldots, \hat{\beta}_{L,p_n})^T$ is a subvector of $\hat{\boldsymbol{\beta}}_L$ and $s_j(\hat{\boldsymbol{\beta}}_L) \in \partial_{b_j} ||\hat{\tilde{\boldsymbol{\beta}}}_L||_1$ is a j-th coefficient of a subgradient of the l_1 norm evaluated at $\hat{\tilde{\boldsymbol{\beta}}}_L$, i.e.:

$$s_j(\hat{\boldsymbol{\beta}}_L) \in \begin{cases} \{\operatorname{sgn} \hat{\boldsymbol{\beta}}_{L,j}\} & \text{if } \hat{\boldsymbol{\beta}}_{L,j} \neq 0 \\ [-1, 1] & \text{if } \hat{\boldsymbol{\beta}}_{L,j} = 0. \end{cases} \tag{4.9}$$

For $\hat{\boldsymbol{\beta}}_{L,-0}$ KKT conditions (4.7) have the same form.

If we denote by:

$$\mathbf{v}(\hat{\boldsymbol{\beta}}_L^T\mathbb{X}) = -\frac{1}{n}\left[\frac{\partial\rho}{\partial b}(\hat{\boldsymbol{\beta}}_L^T\mathbf{X}_1, Y_1), \ldots, \frac{\partial\rho}{\partial b}(\hat{\boldsymbol{\beta}}_L^T\mathbf{X}_n, Y_n)\right]^T,$$
$$\mathbf{s}(\hat{\boldsymbol{\beta}}_L) = [s_1(\hat{\boldsymbol{\beta}}_L), \ldots, s_{p_n}(\hat{\boldsymbol{\beta}}_L)]^T,$$

then (4.7) can be rewritten as:

$$\lambda\mathbf{s}^T(\hat{\boldsymbol{\beta}}_L) = \mathbf{v}^T(\hat{\boldsymbol{\beta}}_L^T\mathbb{X})\tilde{\mathbb{X}}. \tag{4.10}$$

We note that if there exist two solutions of (4.5), namely $\hat{\boldsymbol{\beta}}_L^{(1)}$ and $\hat{\boldsymbol{\beta}}_L^{(2)}$ with $\mathbb{X}\hat{\boldsymbol{\beta}}_L^{(1)} \neq \mathbb{X}\hat{\boldsymbol{\beta}}_L^{(2)}$ and $\rho(\cdot, y)$ is strictly convex, then we would have for $\alpha \in (0,1)$ from strict convexity of ρ:

$$P_n(\alpha\hat{\boldsymbol{\beta}}_L^{(1)} + (1-\alpha)\hat{\boldsymbol{\beta}}_L^{(2)}) < \alpha P_n(\hat{\boldsymbol{\beta}}_L^{(1)}) + (1-\alpha)P_n(\hat{\boldsymbol{\beta}}_L^{(2)}) = P_n(\hat{\boldsymbol{\beta}}_L^{(1)}),$$

what contradicts optimality property of $\hat{\boldsymbol{\beta}}_L^{(1)}$ (proof of strict convexity of P_n is conducted in the same way as in Lemma A.24). Therefore $\mathbb{X}\hat{\boldsymbol{\beta}}_L^{(1)} = \mathbb{X}\hat{\boldsymbol{\beta}}_L^{(2)}$. This means that $R_n(\hat{\boldsymbol{\beta}}_L^{(1)}) = R_n(\hat{\boldsymbol{\beta}}_L^{(2)})$, what gives $||\hat{\tilde{\boldsymbol{\beta}}}_L^{(1)}||_1 = ||\hat{\tilde{\boldsymbol{\beta}}}_L^{(2)}||_1$, if $\lambda > 0$. Moreover, we obtain $\mathbf{s}(\hat{\tilde{\boldsymbol{\beta}}}_L^{(1)}) = \mathbf{s}(\hat{\tilde{\boldsymbol{\beta}}}_L^{(2)})$ in view of (4.10). This implies weak sign consistency for every two Lasso solutions: $\hat{\beta}_{L,j}^{(1)}\hat{\boldsymbol{\beta}}_{L,j}^{(2)} \geq 0$ for $j = 1, \ldots, p_n$, because if $\hat{\beta}_{L,j}^{(1)} > 0$, then from equality $s_j(\hat{\boldsymbol{\beta}}_L^{(2)}) = s_j(\hat{\boldsymbol{\beta}}_L^{(1)}) = 1$ it follows that $\hat{\boldsymbol{\beta}}_{L,j}^{(2)} \geq 0$ (we perform analogous reasoning when $\hat{\beta}_{L,j}^{(1)} < 0$). Note also that the above reasoning shows that if there are two different Lasso solutions, there are uncountably many of them.

Now we will address the question when $\hat{\boldsymbol{\beta}}_L$ is unique. Before we do this, let us introduce the following definition:

Definition 4.1. *We say that $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ has columns in general position, if for every $k < \min\{n, m\}$ no k-dimensional subspace $L \subseteq \mathbb{R}^m$ contains at least $k+2$ points of $\{\pm\boldsymbol{A}^{(1)}, \ldots, \pm\boldsymbol{A}^{(m)}\}$, excluding antipodal pairs.*

From geometric point of view, when $k = 1$, no 3 columns of $\mathbf{A}$ (considered as points in $\mathbb{R}^n$) multiplied by ± 1, say $\pm\mathbf{A}^{(j_1)}, \pm\mathbf{A}^{(j_2)}, \pm\mathbf{A}^{(j_3)}$, can lie on the same line excluding antipodal pairs (i.e. $+\mathbf{A}^{(j)}$ and $-\mathbf{A}^{(j)}$), that is if a line contains points $\pm\mathbf{A}^{(j_1)}$ and $\pm\mathbf{A}^{(j_2)}$ then it may only contain points $\mp\mathbf{A}^{(j_1)}$ and $\mp\mathbf{A}^{(j_2)}$ among $\pm\mathbf{A}^{(j)}$ for $j = 1, \ldots, m$.

Sufficient and necessary conditions for uniqueness of $\hat{\boldsymbol{\beta}}_L$ are known for quadratic loss (see Schneider and Ewald (2017)). Proof of uniqueness provided columns of $\mathbb{X}$ (or $\tilde{\mathbb{X}}$ for the model without intercept) are in general position for quadratic loss can be found in Tibshirani (2013). Moreover, it is noted in Tibshirani and Wasserman (2015), that general position assumption is also sufficient in the case of strictly convex differentiable functions $\rho(\cdot, y)$ provided that $\hat{\boldsymbol{\beta}}_L$ exists. In the case of $p_n + 1 \leq n$ we give sufficient conditions for uniqueness of $\hat{\boldsymbol{\beta}}_L$ (see Lemma A.24) involving strict convexity of $\rho(\cdot, y)$ for all y and a condition on $\mathbb{X}$. In the case of general loss in high-dimensional setup we give sufficient conditions for uniqueness analogous to Rosset et al. (2004) (see also Theorem A.29, which ensures additionally sparsity of Lasso solutions). If $\mathbb{P}_{\tilde{\mathbf{X}}}$ is absolutely continuous distribution with nondegenerate support contained in $\mathbb{R}^{p_n}$ in the sense that Lebesgue's measure $\mu_{p_n}(\operatorname{supp} \tilde{\mathbf{X}}) > 0$, $\rho(\cdot, y)$ is strictly convex and differentiable, then $\hat{\boldsymbol{\beta}}_L$ is unique with probability one (see Theorem A.30). Note that we do not need to impose assumption about distribution of $\mathbb{X}$ to ensure uniqueness of $\hat{\boldsymbol{\beta}}_L$ in any of the proofs in this chapter.

In this chapter we consider cones of the form:

$$C(d, w) = \{\boldsymbol{\Delta} \colon \ ||\boldsymbol{\Delta}_{w^c}||_1 \leq d||\boldsymbol{\Delta}_w||_1\}, \tag{4.11}$$

where $d > 0$, $\boldsymbol{\Delta} \in \mathbb{R}^{p_n+1}$ and $w \subseteq \{0, 1, \ldots, p_n\}$, $w^c = \{1, \ldots, p_n\} \setminus w$ and $\boldsymbol{\Delta}_w = (\Delta_{w_1}, \ldots, \Delta_{w_k})$ for $w = \{w_1, \ldots, w_k\}$. Cones $C(d, w)$ are of special importance, because we prove that $\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^* \in \mathcal{C}$ in the Theorem A.42 for the logistic model with intercept, where

$$\mathcal{C} = C(3, s_0^*), \tag{4.12}$$

$s_0^* = s^* \cup \{0\}$ and $s^* = \{i \in \{1, \ldots, p_n\} : \beta_i^* \neq 0\}$ was defined in (1.4). For the model without intercept the cone is defined analogously but with $\boldsymbol{\Delta} \in \mathbb{R}^{p_n}$ and $w \subseteq \{1, \ldots, p_n\}$. In the cone $C(d, w)$ we define a quantity κ which can be regarded as generalized minimal eigenvalue of a matrix in high-dimensional setup. For the logistic model with intercept we are interested in:

$$\kappa = \inf_{\boldsymbol{\Delta} \in \mathcal{C}} \frac{\boldsymbol{\Delta}^T H(\boldsymbol{\beta}^*) \boldsymbol{\Delta}}{\boldsymbol{\Delta}^T \boldsymbol{\Delta}}, \tag{4.13}$$

$$\kappa_n = \inf_{\boldsymbol{\Delta} \in \mathcal{C}} \frac{\boldsymbol{\Delta}^T H_n(\boldsymbol{\beta}^*) \boldsymbol{\Delta}}{\boldsymbol{\Delta}^T \boldsymbol{\Delta}}, \tag{4.14}$$

where $\mathcal{C}$ was defined in (4.12), $H(\mathbf{b}) = D^2 R(\mathbf{b}) = \mathbb{E}(\mathbf{X}\mathbf{X}^T q_L'(\mathbf{X}_1^T \mathbf{b}))$ is expected value of Hessian. Moreover, empirical Hessian based on all predictors is $H_n(\mathbf{b}) = D^2 R_n(\mathbf{b}) = \frac{\mathbb{X}^T \Pi(\mathbf{b}) \mathbb{X}}{n}$, with $\Pi(\mathbf{b}) = \operatorname{diag}(q_L'(\mathbf{X}_1^T \mathbf{b}), \ldots, q_L'(\mathbf{X}_n^T \mathbf{b}))$. For the model without intercept we define for $\varepsilon > 0$:

$$\kappa_{\mathbf{H}}(\varepsilon) = \inf_{\boldsymbol{\Delta} \in \tilde{\mathcal{C}}_\varepsilon} \frac{\boldsymbol{\Delta}^T \mathbf{H} \boldsymbol{\Delta}}{\boldsymbol{\Delta}^T \boldsymbol{\Delta}}, \tag{4.15}$$

where $\mathbf{H} \in \mathbb{R}^{p_n \times p_n}$ is non-negative definite matrix and

$$\tilde{\mathcal{C}}_\varepsilon = C(3 + \varepsilon, s^*).$$

Moreover, throughout this chapter we introduce a following notation:

$$B_1(r) = \{\boldsymbol{\Delta}\colon\ ||\boldsymbol{\Delta}||_1 \leq r\}, \tag{4.16}$$

$$W(\mathbf{b}) = R(\mathbf{b}) - R(\boldsymbol{\beta}^*), \tag{4.17}$$

$$W_n(\mathbf{b}) = R_n(\mathbf{b}) - R_n(\boldsymbol{\beta}^*), \tag{4.18}$$

$$S(r) = \sup_{\mathbf{b}:\mathbf{b}-\boldsymbol{\beta}^* \in B_1(r)} |W(\mathbf{b}) - W_n(\mathbf{b})|, \tag{4.19}$$

$$\beta^*_{min} = \min_{i \in s^*} |\beta^*_i|. \tag{4.20}$$

We will need the following margin condition for model without intercept in Lemma 4.13 and Theorem 4.15:

(MC) There exist $\vartheta, \varepsilon, \delta > 0$ and non-negative definite matrix $\mathbf{H} \in \mathbb{R}^{p_n \times p_n}$ such that for all $\mathbf{b} \in \Theta$ with $\mathbf{b} - \boldsymbol{\beta}^* \in \tilde{\mathcal{C}}_\varepsilon \cap B_1(\delta)$ we have

$$R(\mathbf{b}) - R(\boldsymbol{\beta}^*) \geq \frac{\vartheta}{2}(\mathbf{b} - \boldsymbol{\beta}^*)^T \mathbf{H} (\mathbf{b} - \boldsymbol{\beta}^*).$$

The above condition can be viewed as a weaker version of strong convexity of function R in the restricted neighbourhood of $\boldsymbol{\beta}^*$ (namely in the intersection of ball $B_1(\delta)$ and cone $\tilde{\mathcal{C}}_\varepsilon$). We stress the fact that $\mathbf{H}$ does not need to be positive definite, as in the Section 4.2 we use (MC) together with stronger conditions than $\kappa_{\mathbf{H}}(\varepsilon) > 0$ - in this situation right hand side of inequality in (MC) is positive. We also do not require here twice-differentiability of R and we note in particular that condition (MC) is satisfied in the case of logistic loss, $\mathbf{X}$ being bounded random variable and $\mathbf{H} = D^2 R(\boldsymbol{\beta}^*)$ - see Lemma A.54. From the Lemma A.55 it also follows that (MC) is satisfied for quadratic loss, $\mathbf{X}$ satisfying $\mathbb{E}||\mathbf{X}||_2^2 < \infty$ and $\mathbf{H} = D^2 R(\boldsymbol{\beta}^*)$. Similar condition to (MC) (called Restricted Strict Convexity) was considered in Negahban et al. (2012) for empirical risk R_n in the context of l_1 regularization:

$$R_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - R_n(\boldsymbol{\beta}^*) \geq D R_n(\boldsymbol{\beta}^*)^T \boldsymbol{\Delta} + \kappa_L ||\boldsymbol{\Delta}||^2 - \tau^2(\boldsymbol{\beta}^*)$$

for all $\boldsymbol{\Delta} \in C(3, s^*)$ and some $\kappa_L > 0$ and tolerance function τ.

Another important assumption, used in the Theorem 4.15 and Lemma 4.14 is the Lipschitz property of ρ :

(LL) $\exists L > 0\ \forall b_1, b_2 \in \mathbb{R}, y \in \{0,1\}\colon\ |\rho(b_1, y) - \rho(b_2, y)| \leq L|b_1 - b_2|$.

4.1. Logistic loss - model with intercept

The main theorem in this section is Theorem 4.9 which is a probabilistic version of Theorem 5 in Fan et al. (2014a) and is also a generalization of this theorem to the case of the model with intercept. Main assumptions of Theorem 5 in Fan et al. (2014a) for the case of deterministic $\mathbb{X}$ are $||DR_n(\boldsymbol{\beta}^*)||_\infty \leq \lambda/2$ and $\max |X_{ij}| \leq \kappa_n/(20\lambda|s^*|)$. Here we will show that these conditions hold with sufficiently high probability, and thus the consistency result for Lasso holds with probability tending to 1 in appropriate setup (see Remark 4.11). We stress that assumption about logistic loss cannot be omitted in Theorem 4.9 as it is essential in the proof of Lemma 4.8. In this section we will denote $s_0^* = s^* \cup \{0\}$ and we assume that unique $\boldsymbol{\beta}^*$ exists.

Lemma 4.2. *(Corollary 8.2 in van de Geer (2016)) Let $Z_1, \ldots, Z_n$ be independent random variables such that for some constant L_0 they satisfy*

$$C_0^2 = \max_{i=1,\ldots,n} \mathbb{E} \exp(|Z_i|/L_0) < \infty.$$

Then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - EZ_i) \geq 2L_0\left(C_0\left(\frac{2t}{n}\right)^{1/2} + \frac{t}{n}\right)\right) \leq e^{-t}.$$

First we prove

Lemma 4.3. *Assume that X_{1j}, $L_0, C_0 > 0$ are such that for all n*

$$\max_{j=0,1,\ldots,p_n} \mathbb{E} \exp\left(\frac{X_{1j}^2}{4L_0}\right) \leq C_0^2. \tag{4.21}$$

Then for any $t > 0$ and $n > t/(2C_0^2)$ we have with probability at least $1 - 2(p_n+1)^2 e^{-t}$ for any $\boldsymbol{\Delta} \in \mathbb{R}^{p_n+1}$:

$$\left|\boldsymbol{\Delta}^T H_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta} - \boldsymbol{\Delta}^T H(\boldsymbol{\beta}^*)\boldsymbol{\Delta}\right| \leq 4||\boldsymbol{\Delta}||_1^2 L_0 C_0 \left(\frac{2t}{n}\right)^{\frac{1}{2}}. \tag{4.22}$$

Proof. Note that

$$\begin{aligned}
|\boldsymbol{\Delta}^T H_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta} - \boldsymbol{\Delta}^T H(\boldsymbol{\beta}^*)\boldsymbol{\Delta}| &= \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i^T\boldsymbol{\Delta})^2 q_L'(\mathbf{X}_i^T\boldsymbol{\beta}^*) - \mathbb{E}(\mathbf{X}_1^T\boldsymbol{\Delta})^2 q_L'(\mathbf{X}_1^T\boldsymbol{\beta}^*)\right| \\
&\leq \sum_{j,k=0}^{p_n} |\Delta_j\Delta_k| \left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik}q_L'(\mathbf{X}_i^T\boldsymbol{\beta}^*) - \mathbb{E}X_{1j}X_{1k}q_L'(\mathbf{X}_1^T\boldsymbol{\beta}^*)\right| \\
&\leq ||\boldsymbol{\Delta}||_1^2 \max_{j,k=0,\ldots,p_n} \left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik}q_L'(\mathbf{X}_i^T\boldsymbol{\beta}^*) - \mathbb{E}X_{1j}X_{1k}q_L'(\mathbf{X}_1^T\boldsymbol{\beta}^*)\right|.
\end{aligned}$$

It follows from Lemma 4.2 applied with $Z_i = X_{ij}X_{ik}q_L'(\mathbf{X}_i^T\boldsymbol{\beta}^*)$ that for any j, k we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik}q_L'(\mathbf{X}_i^T\boldsymbol{\beta}^*) - \mathbb{E}X_{1j}X_{1k}q_L'(\mathbf{X}_1^T\boldsymbol{\beta}^*) \geq 2L_0\left(C_0\left(\frac{2t}{n}\right)^{\frac{1}{2}} + \frac{t}{n}\right)\right) \leq e^{-t}$$

since we have ($|q_L'(s)| \leq 1/4$):

$$\mathbb{E}\exp\left(\frac{|X_{1j}X_{1k}q_L'(\mathbf{X}_1^T\boldsymbol{\beta}^*)|}{L_0}\right) \leq \mathbb{E}\exp\left(\frac{|X_{1j}X_{1k}|}{4L_0}\right) \leq \mathbb{E}\exp\left(\frac{X_{ij}^2 + X_{ik}^2}{8L_0}\right)$$

$$\leq \frac{1}{2}\left(\mathbb{E}\exp\left(\frac{X_{1j}^2}{4L_0}\right)+\mathbb{E}\exp\left(\frac{X_{1k}^2}{4L_0}\right)\right)\leq \frac{1}{2}(C_0^2+C_0^2)=C_0^2.$$

If $n > t/(2C_0^2)$, then it follows that:

$$C_0\left(\frac{2t}{n}\right)^{\frac{1}{2}} > \frac{t}{n}.$$

Now union inequality is used to obtain the result. Note that the set having probability at least $1-2(p_n+1)^2e^{-t}$ on which (4.22) holds can be chosen independently of $\boldsymbol{\Delta}$. □

Corollary 4.4. *If assumptions of Lemma 4.3 are satisfied, then for $t \in (0, 8C_0^2L_0)$ we have:*

$$\mathbb{P}\left(\exists\boldsymbol{\Delta}\in\mathbb{R}^{p_n+1}\colon\ \frac{\left|\boldsymbol{\Delta}^T(H_n(\boldsymbol{\beta}^*)-H(\boldsymbol{\beta}^*))\boldsymbol{\Delta}\right|}{||\boldsymbol{\Delta}||_1^2}>t\right)\leq 2(p_n+1)^2\exp\left(-\frac{t^2n}{32L_0^2C_0^2}\right). \quad (4.23)$$

We define

$$A_1=\{\kappa_n\geq\kappa/2\},\quad A_2=\left\{\max_{\substack{i=1,\ldots,n\\ j=0,1,\ldots,p_n}}|X_{ij}|\leq\frac{\kappa}{40\lambda|s_0^*|}\right\},\quad A_3=\{||DR_n(\boldsymbol{\beta}^*)||_\infty\leq\lambda/2\}.$$

Lemma 4.5. *If assumptions of Lemma 4.3 are satisfied, $\kappa \leq M$ for some $M > 0$, then we have:*

$$\mathbb{P}(A_1)\geq 1-2(p_n+1)^2\exp\left(-\frac{\kappa^2n}{C_1|s_0^*|^2}\right), \quad (4.24)$$

where $C_1 = 32^3L_0^2\tilde{C}_0^2$ and $\tilde{C}_0^2 = \max(C_0^2, M/(256L_0))$.

Proof. Firstly, we observe that if $\boldsymbol{\Delta}\in\mathcal{C}$, then:

$$\frac{||\boldsymbol{\Delta}||_1^2}{||\boldsymbol{\Delta}||_2^2}\leq\frac{(4||\boldsymbol{\Delta}_{s_0^*}||_1)^2}{||\boldsymbol{\Delta}||_2^2}\leq\frac{16|s_0^*|||\boldsymbol{\Delta}_{s_0^*}||_2^2}{||\boldsymbol{\Delta}||_2^2}\leq 16|s_0^*|.$$

Now, if for some $t > 0$ and all $\boldsymbol{\Delta}\in\mathbb{R}^{p_n+1}$ we have

$$|\boldsymbol{\Delta}^TH_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta}-\boldsymbol{\Delta}^TH(\boldsymbol{\beta}^*)\boldsymbol{\Delta}|\leq t||\boldsymbol{\Delta}||_1^2,$$

then we obtain:

$$S=\sup_{\boldsymbol{\Delta}\in\mathcal{C}}\left|\frac{\boldsymbol{\Delta}^TH_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\boldsymbol{\Delta}}-\frac{\boldsymbol{\Delta}^TH(\boldsymbol{\beta}^*)\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\boldsymbol{\Delta}}\right|=\sup_{\boldsymbol{\Delta}\in\mathcal{C}}\frac{||\boldsymbol{\Delta}||_1^2}{||\boldsymbol{\Delta}||_2^2}\frac{\left|\boldsymbol{\Delta}^TH_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta}-\boldsymbol{\Delta}^TH(\boldsymbol{\beta}^*)\boldsymbol{\Delta}\right|}{||\boldsymbol{\Delta}||_1^2}$$
$$\leq 16|s_0^*|\sup_{\boldsymbol{\Delta}\in\mathcal{C}}\frac{\left|\boldsymbol{\Delta}^TH_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta}-\boldsymbol{\Delta}^TH(\boldsymbol{\beta}^*)\boldsymbol{\Delta}\right|}{||\boldsymbol{\Delta}||_1^2}\leq 16|s_0^*|t.$$

On the other hand, observe that for all $\boldsymbol{\Delta}\in\mathcal{C}$ we have:

$$\kappa\leq\frac{\boldsymbol{\Delta}^TH(\boldsymbol{\beta}^*)\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\boldsymbol{\Delta}}\leq\frac{\boldsymbol{\Delta}^TH_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\boldsymbol{\Delta}}+S. \quad (4.25)$$

Taking infimum of right-hand side yields $\kappa \leq \kappa_n + S$. This means that if $S \leq \frac{\kappa}{2}$, then $\kappa_n \geq \frac{\kappa}{2}$. Finally, in view of above inequalities and Corollary 4.4 we get:

$$\mathbb{P}(A_1)\geq\mathbb{P}\left(S\leq\frac{\kappa}{2}\right)\geq\mathbb{P}\left(\forall\boldsymbol{\Delta}\in\mathbb{R}^{p_n+1}\colon\ |\boldsymbol{\Delta}^TH_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta}-\boldsymbol{\Delta}^TH(\boldsymbol{\beta}^*)\boldsymbol{\Delta}|\leq\frac{\kappa||\boldsymbol{\Delta}||_1^2}{32|s_0^*|}\right)$$
$$\geq 1-2(p_n+1)^2\exp\left(-\frac{\kappa^2n}{32^3|s_0^*|^2L_0^2C_0^2}\right).$$

The last inequality in the above chain of inequalities holds, if

$$\frac{\kappa}{32|s_0^*|} \leq 8C_0^2 L_0.$$

If this condition is not satisfied, then we note that it is sufficient to replace C_0 in Lemma 4.3 by a larger constant, therefore the Lemma holds for:

$$\tilde{C}_0^2 = \max\left(C_0^2, \frac{M}{256L_0}\right),$$

as we have:

$$\frac{\kappa}{32|s_0^*|} \leq \frac{M}{32} \leq 8\tilde{C}_0^2 L_0.$$

□

Remark 4.6. *If $X_{ij} \sim Subg(\sigma_{jn}^2)$ then (4.21) is satisfied with $L_0 = s_n^2$ and $C_0 = \sqrt[4]{2}$, what follows from part 5 in Lemma A.32. Hence (4.24) is satisfied with $C_1 = 32^3 s_n^4 \max\{\sqrt{2}, M/(256s_n^2)\}$.*

Lemma 4.7. *If $X_{ij} \sim Subg(\sigma_{jn}^2)$, $\kappa \leq M$ for some $M > 0$ and $\kappa \geq 40|s_0^*|\lambda$, then:*

$$\mathbb{P}(A_1 \cap A_2) \geq 1 - 2(p_n+1)^2 \exp\left(-\frac{\kappa^2 n}{C_1|s_0^*|^2}\right) - 2np_n \exp\left(-\frac{\kappa^2}{C_2\lambda^2|s_0^*|^2}\right),$$

where $C_2 = 3200 s_n^2$ and C_1 was defined in Remark 4.6.

Proof. As $X_{ij} \sim Subg(\sigma_{jn}^2)$, we have in view of union inequality, part 3 in Lemma A.32 and inequality $s_n^2 \geq \sigma_{jn}^2$ for $t \geq 1$:

$$\mathbb{P}\left(\max_{\substack{i=1,\dots,n\\ j=0,1,\dots,p_n}} |X_{ij}| > t\right) = \mathbb{P}\left(\bigcup_{\substack{i=1,\dots,n\\ j=1,\dots,p_n}} \{|X_{ij}| > t\}\right) \leq \sum_{\substack{i=1,\dots,n\\ j=1,\dots,p_n}} \mathbb{P}(|X_{ij}| > t)$$
$$\leq 2np_n \exp\left(-\frac{t^2}{2s_n^2}\right).$$

Thus, we obtain from the above inequality (note that $\kappa/(40\lambda|s_0^*|) > 1$):

$$\mathbb{P}(A_1 \cap A_2^c) = \mathbb{P}\left(\kappa_n \geq \frac{\kappa}{2} \wedge \max_{\substack{i=1,\dots,n\\ j=0,1,\dots,p_n}} |X_{ij}| > \frac{\kappa_n}{20\lambda|s_0^*|}\right) \leq \mathbb{P}\left(\max_{\substack{i=1,\dots,n\\ j=0,1,\dots,p_n}} |X_{ij}| > \frac{\kappa}{40\lambda|s_0^*|}\right)$$
$$\leq 2np_n \exp\left(-\frac{\kappa^2}{3200\lambda^2|s_0^*|^2 s_n^2}\right).$$

The theorem follows from the union inequality: $\mathbb{P}((A_1 \cap A_2)^c) \leq \mathbb{P}(A_1^c) + \mathbb{P}(A_1 \cap A_2^c)$. □

Lemma 4.8. *If $X_{ij} \sim Subg(\sigma_{jn}^2)$, then we have*

$$\mathbb{P}(A_3) \geq 1 - 2p_n \exp\left(-\frac{\lambda^2 n}{8\tau^2 s_n^2}\right) - 2\exp\left(-\frac{\lambda^2 n}{8}\right),$$

where $\tau = e^{\frac{13}{24}} \cdot 4/\sqrt[4]{27} \leq 3.02$.

Proof. Let $Z_i = Y_i - q_L(\mathbf{X}_i^T\boldsymbol{\beta}^*)$. Then we have $|Z_i| \leq 1$ and $\mathbb{E}X_{ij}Z_i = 0$ in view of normal equations 2.11 for $i = 1, \ldots, n$ and $j = 0, 1, \ldots, p_n$. Hence, using Lemmas A.38 and A.33 yields respectively $X_{ij}Z_i \sim Subg(\tau^2\sigma_{jn}^2)$ for $j = 1, \ldots, p_n$ and $X_{ij}Z_i \sim Subg(1)$ for $j = 0$. As observations $(\mathbf{X}_i, Y_i)$ are independent, we have in view of Lemma A.34:

$$\sum_{i=1}^{n} X_{ij}Z_i \sim \begin{cases} Subg(n), & \text{if } j = 0, \\ Subg(n\tau^2\sigma_{jn}^2), & \text{if } j = 1, \ldots, p_n. \end{cases}$$

Thus from the union inequality and part 3 in Lemma A.32 we obtain

$$\begin{aligned}\mathbb{P}(A_3^c) &= \mathbb{P}\left(||DR_n(\boldsymbol{\beta}^*)||_\infty > \frac{\lambda}{2}\right) = \mathbb{P}\left(\bigcup_{j=0}^{p_n}\left\{\sum_{i=1}^{n} X_{ij}Z_i > \frac{n\lambda}{2}\right\}\right) \\ &\leq \sum_{j=0}^{p_n}\mathbb{P}\left(\sum_{i=1}^{n} X_{ij}Z_i > \frac{n\lambda}{2}\right) \leq 2\exp\left(-\frac{n\lambda^2}{8}\right) + 2p_n\exp\left(-\frac{n\lambda^2}{8\tau^2 s_n^2}\right).\end{aligned}$$

□

Theorem 4.9. *If $X_{ij} \sim Subg(\sigma_{jn}^2)$ for $j = 1, \ldots, p_n$, $\kappa \leq M$ for some $M > 0$ and $\kappa \geq 40\lambda|s_0^*|$, then*

$$\begin{aligned}\mathbb{P}\left(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_2 > \frac{10\sqrt{|s_0^*|}\lambda}{\kappa}\right) &\leq 2(p_n+1)^2\exp\left(-\frac{\kappa^2 n}{C_1|s_0^*|^2}\right) + 2np_n\exp\left(-\frac{\kappa^2}{C_2\lambda^2|s_0^*|^2}\right) \\ &\quad + 2\exp\left(-\frac{n\lambda^2}{8}\right) + 2p_n\exp\left(-\frac{n\lambda^2}{8\tau^2 s_n^2}\right),\end{aligned}$$

where $\tau = e^{\frac{13}{24}} \cdot 4/\sqrt[4]{27} \leq 3.02$, $C_1 = 32^3 s_n^4 \max\{\sqrt{2}, M/(256s_n^2)\}$ and $C_2 = 3200s_n^2$.

Proof. On the set $A_2 \cap A_3$ assumptions of Theorem A.42 are satisfied and we have:

$$||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_2 \leq 5|s_0^*|^{1/2}\lambda\kappa_n^{-1}.$$

Thus, on the set $A_1 \cap A_2 \cap A_3$ $\kappa_n \geq \kappa/2$ and we obtain:

$$||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_2 \leq 10|s_0^*|^{1/2}\lambda\kappa^{-1}.$$

This means that:

$$\begin{aligned}\mathbb{P}(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_2 > 10|s_0^*|^{1/2}\lambda\kappa^{-1}) &\leq \mathbb{P}((A_1 \cap A_2 \cap A_3)^c) = \mathbb{P}((A_1 \cap A_2)^c \cup A_3^c) \\ &\leq \mathbb{P}((A_1 \cap A_2)^c) + \mathbb{P}(A_3^c).\end{aligned}$$

This completes the proof in view of Lemmas 4.7 and 4.8. □

Corollary 4.10. *If $X_{ij} \sim Subg(\sigma_{jn}^2)$, $0 < \sigma_{jn} \leq s_n$, $m \leq \kappa \leq M$ for some $m, M > 0$, $\limsup_n s_n < \infty$, $|s_0^*|^2 \log p_n = o(n)$, $\lambda^2|s_0^*|^2\log(np_n) = o(1)$ and $\log p_n = o(n\lambda^2)$, then:*

$$\mathbb{P}\left(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_2 \leq \frac{10\sqrt{|s_0^*|}\lambda}{\kappa}\right) \to 1.$$

If additionally $\beta_{min}^*\kappa \geq 20\sqrt{|s_0^*|}\lambda$ (or $\sqrt{|s_0^*|}\lambda = o(\beta_{min}^*)$), then we have:

$$\mathbb{P}\left(\max_{i \in s^{*c}}|\hat{\beta}_{L,i}| \leq \min_{i \in s^*}|\hat{\beta}_{L,i}|\right) \to 1.$$

Proof. First part is proved in Theorem 4.9. For the second part, we observe that:

$$||\hat{\tilde{\boldsymbol{\beta}}}_L - \tilde{\boldsymbol{\beta}}^*||_2 \leq ||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_2 \leq \frac{10\sqrt{|s_0^*|}\lambda}{\kappa} \leq \frac{\beta^*_{min}}{2}$$

and for the remaining part of proof see proof of Corollary 4.16, as it is almost identical (we replace l_1 norm there with l_2). □

Remark 4.11. *(Separation property)*
Conclusion of Corollary 4.10 holds in high-dimensional setup when $p_n = O(e^{cn^\gamma})$, $|s_0^*| = O(n^\xi)$ *and* $\lambda = C_n\sqrt{\log(p_n)/n}$ *for some* $c > 0$, $\gamma \in (0, 0.5)$, $\xi \in (0, 0.5 - \gamma)$ *and for some sequence* C_n *tending to* ∞ *sufficiently slowly, satisfying* $C_n = O(n^u)$, *where* $u \in (0, 0.5 - \gamma - \xi)$.

4.2. General loss - model without intercept

The main Theorem in this section is Theorem 4.15. Idea of the proof is based on fact that if $S(r)$ defined in (4.19) is sufficiently small, then $\hat{\boldsymbol{\beta}}_L$ lies in a ball $\{\boldsymbol{\Delta} \in \mathbb{R}^{p_n} : \ ||\boldsymbol{\Delta} - \boldsymbol{\beta}^*||_1 \leq r\}$ (see Lemma 4.13). In Lemma 4.14 we prove a tail inequality for $S(r)$, what finally gives us Theorem 4.15.

In this section $\hat{\boldsymbol{\beta}}_L$ stands for $\hat{\boldsymbol{\beta}}_{L,-0}$ defined in (4.6). Quantities $W(\mathbf{v}), W_n(\mathbf{v})$ and $S(r)$ are defined in (4.17) - (4.19) respectively.

Lemma 4.12. *(Basic inequality). Let* $\rho(\cdot, y)$ *be convex function for all* y. *If*

$$u = \frac{r}{r + ||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}||_1}, \quad \boldsymbol{v} = u\hat{\boldsymbol{\beta}}_L + (1-u)\boldsymbol{\beta}^*,$$

then:

$$W(\boldsymbol{v}) + \lambda||\boldsymbol{v} - \boldsymbol{\beta}^*||_1 \leq S(r) + 2\lambda||\boldsymbol{v}_{s^*} - \boldsymbol{\beta}^*_{s^*}||_1.$$

Proof. Firstly, observe that from convexity of ρ function R_n is convex. Moreover, from the definition of $\hat{\boldsymbol{\beta}}_L$ we get the inequality:

$$W_n(\hat{\boldsymbol{\beta}}_L) = R_n(\hat{\boldsymbol{\beta}}_L) - R_n(\boldsymbol{\beta}^*) \leq \lambda(||\boldsymbol{\beta}^*||_1 - ||\hat{\boldsymbol{\beta}}_L||_1). \tag{4.26}$$

We note that $\mathbf{v} - \boldsymbol{\beta}^* \in B_1(r)$, as we have:

$$||\mathbf{v} - \boldsymbol{\beta}^*||_1 = \frac{||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1}{r + ||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1} \cdot r \leq r. \tag{4.27}$$

By definition of W_n, convexity of R_n, (4.27) and definition of S we have:

$$\begin{aligned} W(\mathbf{v}) &= W(\mathbf{v}) - W_n(\mathbf{v}) + R_n(\mathbf{v}) - R_n(\boldsymbol{\beta}^*) \\ &\leq W(\mathbf{v}) - W_n(\mathbf{v}) + u(R_n(\hat{\boldsymbol{\beta}}_L) - R_n(\boldsymbol{\beta}^*)) \leq S(r) + uW_n(\hat{\boldsymbol{\beta}}_L). \end{aligned} \tag{4.28}$$

From the convexity of l_1 norm, (4.28), (4.26), $||\boldsymbol{\beta}^*||_1 = ||\boldsymbol{\beta}^*_{s^*}||_1$ and triangle inequality it follows that:

$$W(\mathbf{v}) + \lambda||\mathbf{v}||_1 \leq W(\mathbf{v}) + \lambda u||\hat{\boldsymbol{\beta}}_L||_1 + \lambda(1-u)||\boldsymbol{\beta}^*||_1$$

$$\leq S(r) + uW_n(\hat{\boldsymbol{\beta}}_L) + u\lambda(||\hat{\boldsymbol{\beta}}_L||_1 - ||\boldsymbol{\beta}^*||_1) + \lambda||\boldsymbol{\beta}^*||_1$$
$$\leq S(r) + \lambda||\boldsymbol{\beta}^*||_1 \leq S(r) + \lambda||\boldsymbol{\beta}^* - \mathbf{v}_{s^*}||_1 + \lambda||\mathbf{v}_{s^*}||_1. \tag{4.29}$$

Hence:

$$W(\mathbf{v}) + \lambda||\mathbf{v} - \boldsymbol{\beta}^*||_1 = (W(\mathbf{v}) + \lambda||\mathbf{v}||_1) + \lambda(||\mathbf{v} - \boldsymbol{\beta}^*||_1 - ||\mathbf{v}||_1)$$
$$\leq S(r) + \lambda||\boldsymbol{\beta}^* - \mathbf{v}_{s^*}||_1 + \lambda||\mathbf{v}_{s^*}||_1 + \lambda(||\mathbf{v} - \boldsymbol{\beta}^*||_1 - ||\mathbf{v}||_1) = S(r) + 2\lambda||\boldsymbol{\beta}^* - \mathbf{v}_{s^*}||_1.$$

□

Lemma 4.13. *Let $\rho(\cdot, y)$ be convex function for all y. Assume that $\lambda > 0$. Moreover, assume margin condition (MC) with constants $\vartheta, \epsilon, \delta > 0$ and some non-negative definite matrix $\boldsymbol{H} \in \mathbb{R}^{p_n \times p_n}$. If for some $r \in (0, \delta]$ we have $S(r) \leq \bar{C}\lambda r$ and $2|s^*|\lambda \leq \kappa_{\boldsymbol{H}}(\varepsilon)\vartheta\tilde{C}r$, where $\bar{C} = \varepsilon/(8 + 2\varepsilon)$ and $\tilde{C} = 2/(4 + \varepsilon)$, then*

$$||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 \leq r.$$

Proof. Let u and $\mathbf{v}$ be defined as in Lemma 4.12. Observe that $||\mathbf{v} - \boldsymbol{\beta}^*||_1 \leq r/2$ is equivalent to $||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 \leq r$, as the function $f(x) = rx/(x + r)$ is increasing, $f(r) = r/2$ and $f(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1) = ||\mathbf{v} - \boldsymbol{\beta}^*||_1$. Let $C = 1/(4 + \varepsilon)$. We consider two cases:
(i) $||\mathbf{v}_{s^*} - \boldsymbol{\beta}^*_{s^*}||_1 \leq Cr$:
From the basic inequality (Lemma 4.12) we have:

$$||\mathbf{v} - \boldsymbol{\beta}^*||_1 \leq \lambda^{-1}(W(\mathbf{v}) + \lambda||\mathbf{v} - \boldsymbol{\beta}^*||_1) \leq \lambda^{-1}S(r) + 2||\mathbf{v}_{s^*} - \boldsymbol{\beta}^*_{s^*}||_1 \leq \bar{C}r + 2Cr = \frac{r}{2}.$$

(ii) $||\mathbf{v}_{s^*} - \boldsymbol{\beta}^*_{s^*}||_1 > Cr$:
Firstly, we observe, that $||\mathbf{v}_{s^{*c}}||_1 < (1 - C)r$, otherwise we would have $||\mathbf{v} - \boldsymbol{\beta}^*||_1 > r$ which contradicts (4.27).
Now, we observe that $\mathbf{v} - \boldsymbol{\beta}^* \in \tilde{\mathcal{C}}_\varepsilon$, as we have from definition of C and assumption of this case:

$$||\mathbf{v}_{s^{*c}}||_1 < (1 - C)r = (3 + \varepsilon)Cr < (3 + \varepsilon)||\mathbf{v}_{s^*} - \boldsymbol{\beta}^*_{s^*}||_1.$$

By inequality between l_1 and l_2 norm, definition of $\kappa_{\mathbf{H}}(\varepsilon)$, inequality $ca^2/4 + b^2/c \geq ab$ and margin condition (MC) (which holds because $\mathbf{v} - \boldsymbol{\beta}^* \in B_1(r) \subseteq B_1(\delta)$ from (4.27)) it may be concluded that:

$$||\mathbf{v}_{s^*} - \boldsymbol{\beta}^*_{s^*}||_1 \leq \sqrt{|s^*|}||\mathbf{v}_{s^*} - \boldsymbol{\beta}^*_{s^*}||_2 \leq \sqrt{|s^*|}||\mathbf{v} - \boldsymbol{\beta}^*||_2 \leq \sqrt{|s^*|}\sqrt{\frac{(\mathbf{v} - \boldsymbol{\beta}^*)^T\mathbf{H}(\mathbf{v} - \boldsymbol{\beta}^*)}{\kappa_{\mathbf{H}}(\varepsilon)}}$$
$$\leq \frac{\vartheta(\mathbf{v} - \boldsymbol{\beta}^*)^T\mathbf{H}(\mathbf{v} - \boldsymbol{\beta}^*)}{4\lambda} + \frac{|s^*|\lambda}{\vartheta\kappa_{\mathbf{H}}(\varepsilon)} \leq \frac{W(\mathbf{v})}{2\lambda} + \frac{|s^*|\lambda}{\vartheta\kappa_{\mathbf{H}}(\varepsilon)}. \tag{4.30}$$

Hence from the basic inequality (Lemma 4.12) and above inequality it follows that:

$$W(\mathbf{v}) + \lambda||\mathbf{v} - \boldsymbol{\beta}^*||_1 \leq S(r) + 2\lambda||\mathbf{v}_{s^*} - \boldsymbol{\beta}^*_{s^*}||_1 \leq S(r) + W(\mathbf{v}) + \frac{2|s^*|\lambda^2}{\vartheta\kappa_{\mathbf{H}}(\varepsilon)}.$$

Substracting $W(\mathbf{v})$ from both sides of above inequality, using assumption about S, inequality about $|s^*|$ and definition of $\tilde{C}$ yields:

$$||\mathbf{v} - \boldsymbol{\beta}^*||_1 \leq \frac{S(r)}{\lambda} + \frac{2|s^*|\lambda}{\vartheta\kappa_{\mathbf{H}}(\varepsilon)} \leq \bar{C}r + \frac{2|s^*|\lambda}{\vartheta\kappa_{\mathbf{H}}(\varepsilon)} \leq (\bar{C} + \tilde{C})r = \frac{r}{2}.$$

□

Lemma 4.14. *Let $\rho(\cdot, y)$ be convex function for all y and satisfy Lipschitz condition (LL) for all b_1, b_2, y. Assume that X_{ij} for $j \geq 1$ are subgaussian $Subg(\sigma_{jn}^2)$ where $\sigma_{jn} \leq s_n$. Then for $r, t > 0$:*

$$\mathbb{P}(S(r) > t) \leq \frac{14Lrs_n\sqrt{\log(p_n \vee 2)}}{t\sqrt{n}}.$$

Proof. From the Chebyshev inequality (first inequality), symmetrization inequality (Theorem A.40) and Talagrand - Ledoux inequality (Theorem A.41) we have for $t > 0$ (where $(\varepsilon_i)_{i=1,\ldots,n}$ are Rademacher variables independent of $\tilde{\mathbb{X}}$):

$$\begin{aligned}\mathbb{P}(S(r) > t) &\leq \frac{\mathbb{E}S(r)}{t} \leq \frac{2}{t}\mathbb{E}\sup_{\mathbf{b}\in\mathbb{R}^{p_n}:\mathbf{b}-\boldsymbol{\beta}^*\in B_1(r)}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\rho(\mathbf{X}_i^T\mathbf{b}, Y_i) - \rho(\mathbf{X}_i^T\boldsymbol{\beta}^*, Y_i))\right| \\ &\leq \frac{4L}{t}\mathbb{E}\sup_{\mathbf{b}\in\mathbb{R}^{p_n}:\mathbf{b}-\boldsymbol{\beta}^*\in B_1(r)}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_i^T(\mathbf{b}-\boldsymbol{\beta}^*)\right| \leq \frac{4Lr}{t}\mathbb{E}\max_{j\in\{1,\ldots,p_n\}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_{ij}\right|.\end{aligned}$$

In view of Lemma A.39 we obtain $\varepsilon_i X_{ij} \sim Subg(\sigma_{jn}^2)$. From independence of $(\varepsilon_i\mathbf{X}_i)_{i=1,\ldots,n}$ and Lemma A.34 we have that $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_{ij} \sim Subg(\frac{\sigma_{jn}^2}{n})$. Thus $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_{ij} \sim Subg(\frac{s_n^2}{n})$.
In view of Lemma A.36 we obtain:

$$\mathbb{E}\max_{j\in\{1,\ldots,p_n\}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_{ij}\right| \leq \frac{7}{2}s_n\sqrt{\frac{\log(p_n \vee 2)}{n}}.$$

This ends the proof. □

Theorem 4.15. *Let $\rho(\cdot, y)$ be convex function for all y and satisfy Lipschitz condition (LL). Assume that $X_{ij} \sim Subg(\sigma_{jn}^2)$ with $\sigma_{jn} \leq s_n$, $\boldsymbol{\beta}^*$ defined in (1.8) exists and is unique, margin condition (MC) is satisfied for $\varepsilon, \delta, \vartheta > 0$, non-negative definite matrix $\boldsymbol{H} \in \mathbb{R}^{p_n\times p_n}$ and let*

$$\frac{2|s^*|\lambda}{\vartheta\kappa_{\boldsymbol{H}}(\varepsilon)} \leq \tilde{C}\min\left\{\frac{\beta^*_{min}}{2}, \delta\right\},$$

where $\tilde{C} = 2/(4+\varepsilon)$. Then:

$$\mathbb{P}\left(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 \leq \frac{\beta^*_{min}}{2}\right) \geq 1 - \frac{14(8+2\varepsilon)Ls_n\sqrt{\log(p_n \vee 2)}}{\varepsilon\lambda\sqrt{n}}.$$

Proof. Let:

$$m = \min\left\{\frac{\beta^*_{min}}{2}, \delta\right\}.$$

Lemmas 4.13 and 4.14 imply that:

$$\begin{aligned}\mathbb{P}\left(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 > \frac{\beta^*_{min}}{2}\right) &\leq \mathbb{P}\left(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 > m\right) \leq \mathbb{P}\left(S(m) > \bar{C}\lambda m\right) \\ &\leq \frac{14(8+2\varepsilon)Ls_n\sqrt{\log(p_n \vee 2)}}{\varepsilon\lambda\sqrt{n}}.\end{aligned}$$

□

Corollary 4.16. *(Separation property) If assumptions of Theorem 4.15 are satisfied,* $\log p_n = o(\lambda^2 n)$ *and* $\kappa_{\boldsymbol{H}}(\varepsilon) > d$ *for some* $d, \varepsilon > 0$ *for large* n, $|s^*|\lambda = o(\min\{\beta^*_{min}, 1\})$, *then*

$$\mathbb{P}\left(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 \leq \frac{\beta^*_{min}}{2}\right) \to 1.$$

Moreover

$$\mathbb{P}\left(\max_{i \in s^{*c}} |\hat{\beta}_{L,i}| \leq \min_{i \in s^*} |\hat{\beta}_{L,i}|\right) \to 1.$$

Proof. First part of the corollary follows directly from Theorem 4.15. Now we prove that condition $||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 \leq \beta^*_{min}/2$ implies separation property $\max_{i \in s^{*c}} |\hat{\beta}_{L,i}| \leq \min_{i \in s^*} |\hat{\beta}_{L,i}|$.

Observe that for all $j \in \{1, \ldots, p_n\}$ we have:

$$\frac{\beta^*_{min}}{2} \geq ||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 \geq |\hat{\beta}_{L,j} - \beta^*_j|. \tag{4.31}$$

If $j \in s^*$, then using triangle inequality yields:

$$|\hat{\beta}_{L,j} - \beta^*_j| \geq |\beta^*_j| - |\hat{\beta}_{L,j}| \geq \beta^*_{min} - |\hat{\beta}_{L,j}|.$$

Hence from the above inequality and (4.31) we obtain for $j \in s^*$:

$$|\hat{\beta}_{L,j}| \geq \frac{\beta^*_{min}}{2}.$$

If $j \in s^{*c}$, then $\beta^*_j = 0$ and (4.31) takes the form:

$$|\hat{\beta}_{L,j}| \leq \frac{\beta^*_{min}}{2}.$$

This ends the proof. □

Chapter 5

GIC minimization

Consider an arbitrary family $\mathcal{M} \subseteq 2^{\{1,\dots,p_n\}}$ of models (which may be data-dependent) such that $s^* \in \mathcal{M}, \forall w \in \mathcal{M}: \ |w| \leq k_n$ a.e. and $k_n \in \mathbb{N}_+$ is some sequence. We define Generalized Information Criterion (GIC) as:

$$GIC(w) = nR_n(\hat{\boldsymbol{\beta}}(w)) + a_n|w|, \tag{5.1}$$

where

$$\hat{\boldsymbol{\beta}}(w) = \underset{\mathbf{b}\in\mathbb{R}^{p_n}:\ \mathbf{b}_{w^c}=\mathbf{0}_{|w^c|}}{\arg\min} R_n(\mathbf{b})$$

and $a_n > 0$ is some penalty. Typical examples of a_n include:

- AIC (Akaike Information Criterion): $a_n = 2$,
- BIC (Bayesian Information Criterion): $a_n = \log n$,
- EBIC(d) (Extended BIC): $a_n = \log n + 2d \log p_n$, where $d > 0$.

In this chapter we consider only the model without intercept as the results for GIC minimization generalise easily to the case of the model with intercept. Moreover, throughout the chapter we introduce a following notation:

$$B_2(r) = \{\boldsymbol{\Delta} \in \mathbb{R}^{p_n}: \ ||\boldsymbol{\Delta}||_2 \leq r\}, \tag{5.2}$$

$$D_1 = \{\mathbf{b} \in \mathbb{R}^{p_n}: \ \exists w \in \mathcal{M}: \ |w| \leq k_n \wedge s^* \subset w \wedge \operatorname{supp} \mathbf{b} \subseteq w\}, \tag{5.3}$$

$$S_1(r) = \sup_{\mathbf{b}\in D_1:\mathbf{b}-\boldsymbol{\beta}^*\in B_2(r)} |(R_n(\mathbf{b}) - R_n(\boldsymbol{\beta}^*)) - (R(\mathbf{b}) - R(\boldsymbol{\beta}^*))|, \tag{5.4}$$

$$D_2 = \{\mathbf{b} \in \mathbb{R}^{p_n}: \ \operatorname{supp} \mathbf{b} \subset s^*\}, \tag{5.5}$$

$$S_2(r) = \sup_{\mathbf{b}\in D_2:\mathbf{b}-\boldsymbol{\beta}^*\in B_2(r)} |(R_n(\mathbf{b}) - R_n(\boldsymbol{\beta}^*)) - (R(\mathbf{b}) - R(\boldsymbol{\beta}^*))|. \tag{5.6}$$

We note that such definitions of D_i for $i = 1, 2$ guarantee that if $\mathbf{b} \in D_i$, then $|\operatorname{supp}(\mathbf{b} - \boldsymbol{\beta}^*)| \leq k_n$, what we exploit in Lemma 5.1.

We assume that X_{ij} are subgaussian: $X_{ij} \sim Subg(\sigma_{jn}^2)$, where: $\sigma_{jn} > 0$ and moreover $\limsup_n s_n = \gamma < \infty$, where $s_n = \max_j \sigma_{jn}$.

Moreover, in Section 5.1 we consider the following condition for $\epsilon > 0$, $w \subseteq \{1, \dots, p_n\}$ and some $\theta > 0$:

$C_\epsilon(w)$: $R(\mathbf{b}) - R(\boldsymbol{\beta}^*) \geq \theta||\mathbf{b}-\boldsymbol{\beta}^*||_2^2$ for all $\mathbf{b} \in \mathbb{R}^{p_n}$ such that $\operatorname{supp}\mathbf{b} \subseteq w$ and $\mathbf{b}-\boldsymbol{\beta}^* \in B_2(\epsilon)$.

We note in particular that in view of Lemma A.54 the above condition is satisfied for logistic loss, $\mathbf{X} \in \mathbb{R}^{p_n}$ being bounded, sufficiently small $\epsilon > 0$ and when the following condition is fulfilled:

$$\liminf_n \inf_{\boldsymbol{\Delta}\in\mathbb{R}^{p_n}} \frac{\boldsymbol{\Delta}^T D^2 R(\boldsymbol{\beta}^*)\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\boldsymbol{\Delta}} > 0. \tag{5.7}$$

Moreover, $C_\epsilon(w)$ is satisfied for quadratic loss and $\mathbf{X} \in \mathbb{R}^{p_n}$ satisfying $\mathbb{E}||\mathbf{X}||_2^2 < \infty$ and (5.7) in view of Lemma A.55. Condition (5.7) in the proofs can be replaced also by the following weaker one to guarantee that $C_\epsilon(w)$ is fulfilled:

$$\liminf_n \inf_{\boldsymbol{\Delta}\in\mathbb{R}^{p_n}\,:\,||\boldsymbol{\Delta}||_0\leq k_n} \frac{\boldsymbol{\Delta}^T D^2 R(\boldsymbol{\beta}^*)\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\boldsymbol{\Delta}} > 0. \tag{5.8}$$

We observe also that the conditions (MC) and $C_\epsilon(w)$ are not equivalent, as they hold for $\mathbf{v} = \mathbf{b}-\boldsymbol{\beta}^*$ belonging to different sets: $B_1(r)\cap\tilde{\mathcal{C}}_\varepsilon$ and $B_2(\epsilon)\cap\{\boldsymbol{\Delta}\in\mathbb{R}^{p_n}:\ \operatorname{supp}\boldsymbol{\Delta}\subseteq w\}$ respectively. We note that if the following condition is satisfied for matrix $\mathbf{H}$ in condition (MC):

$$\inf_{\boldsymbol{\Delta}\in\mathbb{R}^{p_n}} \frac{\boldsymbol{\Delta}^T\mathbf{H}\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\boldsymbol{\Delta}} = \lambda_{min} > 0,$$

and (MC) holds for $\mathbf{b}-\boldsymbol{\beta}^* \in B_1(r)$ (instead of for $\mathbf{b}-\boldsymbol{\beta}^* \in \tilde{\mathcal{C}}_\varepsilon \cap B_1(r)$) then we have for $\mathbf{b}-\boldsymbol{\beta}^* \in B_2(r/\sqrt{p_n}) \subseteq B_1(r)$:

$$R(\mathbf{b}) - R(\boldsymbol{\beta}^*) \geq \frac{\vartheta}{2}(\mathbf{b}-\boldsymbol{\beta}^*)^T\mathbf{H}(\mathbf{b}-\boldsymbol{\beta}^*) \geq \frac{\vartheta\lambda_{min}}{2}||\mathbf{b}-\boldsymbol{\beta}^*||_2^2.$$

Furthermore, if

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^{p_n}} \frac{\boldsymbol{\Delta}^T\mathbf{H}\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\boldsymbol{\Delta}} = \lambda_{max},$$

and $C_\epsilon(w)$ holds for all $\mathbf{v} = \mathbf{b}-\boldsymbol{\beta}^* \in B_2(r)$ without restriction on $\operatorname{supp}\mathbf{b}$, then we have for $\mathbf{b}-\boldsymbol{\beta}^* \in B_1(r) \subseteq B_2(r)$:

$$R(\mathbf{b}) - R(\boldsymbol{\beta}^*) \geq \theta||\mathbf{b}-\boldsymbol{\beta}^*||_2^2 \geq \frac{\theta}{\lambda_{max}}(\mathbf{b}-\boldsymbol{\beta}^*)^T\mathbf{H}(\mathbf{b}-\boldsymbol{\beta}^*).$$

Similar condition to $C_\epsilon(w)$ for empirical risk R_n was considered in (Kim and Jeon, 2016, (2.1)) in the context of GIC minimization.

It turns out that condition $C_\epsilon(w)$ together with $\rho(\cdot, y)$ being convex for all y and satisfying Lipschitz condition (LL) are sufficient to establish bounds which ensure GIC consistency for $k_n \ln p_n = o(n)$ and $k_n \ln p_n = o(a_n)$ (see Corollaries 5.3 and 5.5).

Theorems 5.2 and 5.4 state probability inequalities related to GIC consistency respectively on supersets of s^* and on subsets of s^*. Corollaries 5.3 and 5.5 present asymptotic conditions for GIC consistency in the aforementioned situations. Corollaries 5.6 and 5.7 gather conclusions of Corollaries 4.10, 4.16, 5.3 and 5.5 to show consistency of SS procedure (see Pokarowski and Mielniczuk (2015)) in case of subgaussian variables.

5.1. GIC consistency

Lemma 5.1 is similar to Lemma 4.14. However, we bound $S_1(r)$ and $S_2(r)$ on ball $B_2(r)$ instead of $B_1(r)$, which was considered in Lemma 4.14.

Lemma 5.1. *If $X_{ij} \sim Subg(\sigma_{jn}^2)$, $\rho(\cdot, y)$ is Lipschitz with constant $L > 0$ for all y, $r, t > 0$, then:*

1. $\mathbb{P}(S_1(r) \geq t) \leq \frac{14Lr}{\sqrt{nt}}\sqrt{k_n}s_n\sqrt{\ln(p_n \vee 2)}$,
2. $\mathbb{P}(S_2(r) \geq t) \leq \frac{4Lr}{\sqrt{nt}}\sqrt{|s^*|}s_n$.

Proof. Using respectively: Markov's inequality, Lemmas A.40, A.41, Schwarz's inequality, inequality $||\mathbf{v}||_2 \leq \sqrt{||\mathbf{v}||_0}||\mathbf{v}||_\infty$, inequality $||\mathbf{v}_\pi||_\infty \leq ||\mathbf{v}||_\infty$ for $\pi \subseteq \{1, \ldots, p_n\}$ and Lemma A.36 yields:

$$\begin{aligned}
\mathbb{P}(S_1(r) \geq t) &\leq \frac{\mathbb{E}S_1(r)}{t} \leq \frac{2}{nt}\mathbb{E}\sup_{\mathbf{b}\in D_1\colon \mathbf{b}-\boldsymbol{\beta}^*\in B_2(r)}\left|\sum_{i=1}^n \varepsilon_i(\rho(\mathbf{X}_i^T\mathbf{b}, Y_i) - \rho(\mathbf{X}_i^T\boldsymbol{\beta}^*, Y_i))\right| \\
&\leq \frac{4L}{nt}\mathbb{E}\sup_{\mathbf{b}\in D_1\colon \mathbf{b}-\boldsymbol{\beta}^*\in B_2(r)}\left|\sum_{i=1}^n \varepsilon_i\mathbf{X}_i^T(\mathbf{b}-\boldsymbol{\beta}^*)\right| \leq \frac{4Lr}{nt}\mathbb{E}\max_{\pi\subseteq\{1,\ldots,p_n\},|\pi|\leq k_n}\left|\left|\sum_{i=1}^n \varepsilon_i\mathbf{X}_{i,\pi}\right|\right|_2 \\
&\leq \frac{4Lr}{nt}\mathbb{E}\max_{\pi\subseteq\{1,\ldots,p_n\},|\pi|\leq k_n}\sqrt{|\pi|}\left|\left|\sum_{i=1}^n \varepsilon_i\mathbf{X}_{i,\pi}\right|\right|_\infty \leq \frac{4Lr\sqrt{k_n}}{nt}\mathbb{E}\left|\left|\sum_{i=1}^n \varepsilon_i\mathbf{X}_i\right|\right|_\infty \\
&\leq \frac{14Lr}{t\sqrt{n}}\sqrt{k_n}s_n\sqrt{\ln(p_n \vee 2)}.
\end{aligned}$$

Similarly for $S_2(r)$, using inequality $||\mathbf{v}_\pi||_2 \leq ||\mathbf{v}_{s^*}||_2$ which is valid for $\pi \subseteq s^*$, definition of l_2 norm, inequality $\mathbb{E}|Z| \leq \sqrt{\mathbb{E}Z^2}$ and Lemma A.32 p. 2, we obtain:

$$\begin{aligned}
\mathbb{P}(S_2(r) \geq t) &\leq \frac{\mathbb{E}S_2(r)}{t} \leq \frac{2}{nt}\mathbb{E}\sup_{\mathbf{b}\in D_2\colon \mathbf{b}-\boldsymbol{\beta}^*\in B_2(r)}\left|\sum_{i=1}^n \varepsilon_i(\rho(\mathbf{X}_i^T\mathbf{b}, Y_i) - \rho(\mathbf{X}_i^T\boldsymbol{\beta}^*, Y_i))\right| \\
&\leq \frac{4L}{nt}\mathbb{E}\sup_{\mathbf{b}\in D_2\colon \mathbf{b}-\boldsymbol{\beta}^*\in B_2(r)}\left|\sum_{i=1}^n \varepsilon_i\mathbf{X}_i^T(\mathbf{b}-\boldsymbol{\beta}^*)\right| \leq \frac{4Lr}{nt}\mathbb{E}\max_{\pi\subseteq s^*}\left|\left|\sum_{i=1}^n \varepsilon_i\mathbf{X}_{i,\pi}\right|\right|_2 \\
&\leq \frac{4Lr}{nt}\mathbb{E}\left|\left|\sum_{i=1}^n \varepsilon_i\mathbf{X}_{i,s^*}\right|\right|_2 \leq \frac{4Lr}{nt}\sqrt{\mathbb{E}\left|\left|\sum_{i=1}^n \varepsilon_i\mathbf{X}_{i,s^*}\right|\right|_2^2} = \frac{4Lr}{nt}\sqrt{\sum_{j\in s^*}\mathbb{E}\left(\sum_{i=1}^n \varepsilon_i X_{ij}\right)^2} \\
&\leq \frac{4Lr}{\sqrt{nt}}\sqrt{|s^*|}s_n.
\end{aligned}$$

□

Theorem below provides conditions for GIC consistency on supersets of s^* in Corollary 5.3. We note that bound in (5.9) is optimized for

$$r = \sqrt{\frac{4a_n}{\theta n}},$$

however this requires assumption $a_n = o(n)$, as $r < \epsilon$ and consistency on supersets of s^* is obtained when $k_n \ln p_n = o(a_n)$. Corollary 5.3 gives weaker assumptions for a special choice of $r = r_n \to 0$.

Theorem 5.2. *Assume that $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma_{jn}^2)$, condition $C_\epsilon(w)$ holds for some $\epsilon, \theta > 0$ and for every $w \subseteq \{1, \ldots, p_n\}$ such that $|w| \leq k_n$. Then for any $r < \epsilon$ we have:*

$$\mathbb{P}(\min_{w \in \mathcal{M}: s^* \subset w} GIC(w) \leq GIC(s^*)) \leq \frac{14L\sqrt{k_n} s_n \sqrt{\ln(p_n \vee 2)}}{\sqrt{n}} \left(\frac{rn}{a_n} + \frac{4}{\theta r}\right). \tag{5.9}$$

Proof. If $s^* \subset w \in \mathcal{M}$ and $\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^* \in B_2(r)$ then in view of inequalities $R_n(\hat{\boldsymbol{\beta}}(s^*)) \leq R_n(\boldsymbol{\beta}^*)$ and $R(\boldsymbol{\beta}^*) \leq R(\mathbf{b})$ we observe that:

$$\begin{aligned} R_n(\hat{\boldsymbol{\beta}}(s^*)) - R_n(\hat{\boldsymbol{\beta}}(w)) &\leq \sup_{\mathbf{b} \in D_1 \colon \mathbf{b} - \boldsymbol{\beta}^* \in B_2(r)} (R_n(\boldsymbol{\beta}^*) - R_n(\mathbf{b})) \\ &\leq \sup_{\mathbf{b} \in D_1 \colon \mathbf{b} - \boldsymbol{\beta}^* \in B_2(r)} ((R_n(\boldsymbol{\beta}^*) - R(\boldsymbol{\beta}^*)) - (R_n(\mathbf{b}) - R(\mathbf{b}))) \\ &\leq \sup_{\mathbf{b} \in D_1 \colon \mathbf{b} - \boldsymbol{\beta}^* \in B_2(r)} |R_n(\mathbf{b}) - R(\mathbf{b}) - (R_n(\boldsymbol{\beta}^*) - R(\boldsymbol{\beta}^*))| = S_1(r). \end{aligned}$$

Moreover, we observe that: $a_n(|w| - |s^*|) \geq a_n$. Hence, if we have for some $w \supset s^*$: $GIC(w) \leq GIC(s^*)$ then we obtain $nR_n(\hat{\boldsymbol{\beta}}(s^*)) - nR_n(\hat{\boldsymbol{\beta}}(w))) \geq a_n(|w| - |s^*|)$ and from the above inequality we have $S_1(r) \geq \frac{a_n}{n}$. Furthermore, if $\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^* \in B_2(r)^c$ and $r < \epsilon$, then we take:

$$\mathbf{v} = u\hat{\boldsymbol{\beta}}(w) + (1 - u)\boldsymbol{\beta}^*,$$

where $u = r/(r + ||\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^*||_2)$. This means that:

$$||\mathbf{v} - \boldsymbol{\beta}^*||_2 = u||\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^*||_2 = r \cdot \frac{||\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^*||_2}{r + ||\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^*||_2} \geq \frac{r}{2},$$

as function $x/(x + r)$ is increasing with respect to x for $x > 0$. Moreover, we have $||\mathbf{v} - \boldsymbol{\beta}^*||_2 \leq r < \epsilon$. Hence, in view of $C_\epsilon(w)$ condition we get:

$$R(\mathbf{v}) - R(\boldsymbol{\beta}^*) \geq \theta||\mathbf{v} - \boldsymbol{\beta}^*||_2^2 \geq \frac{\theta r^2}{4}.$$

From convexity of R_n we have:

$$R_n(\mathbf{v}) \leq u(R_n(\hat{\boldsymbol{\beta}}(w)) - R_n(\boldsymbol{\beta}^*)) + R_n(\boldsymbol{\beta}^*) \leq R_n(\boldsymbol{\beta}^*).$$

We observe that $\operatorname{supp} \mathbf{v} \subseteq \operatorname{supp} \hat{\boldsymbol{\beta}}(w) \cup \operatorname{supp} \boldsymbol{\beta}^* \subseteq w$, hence $\mathbf{v} \in D_1$. Finally, we have:

$$S_1(r) \geq R_n(\boldsymbol{\beta}^*) - R(\boldsymbol{\beta}^*) - (R_n(\mathbf{v}) - R(\mathbf{v})) \geq R(\mathbf{v}) - R(\boldsymbol{\beta}^*) \geq \frac{\theta r^2}{4}.$$

Hence we obtain the following sequence of inequalities:

$$\begin{aligned} \mathbb{P}(\min_{w \in \mathcal{M}: s^* \subset w} GIC(w) \leq GIC(s^*)) &\leq \mathbb{P}(S_1(r) \geq \frac{a_n}{n}, \forall w \in \mathcal{M} \colon \ \hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^* \in B_2(r)) \\ + \mathbb{P}(\exists w \in \mathcal{M} : s^* \subset w \wedge \hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^* \in B_2(r)^c) &\leq \mathbb{P}(S_1(r) \geq \frac{a_n}{n}) + \mathbb{P}(S_1(r) \geq \frac{\theta r^2}{4}) \\ &\leq \frac{14Lr\sqrt{n}}{a_n} \sqrt{k_n} s_n \sqrt{\ln(p_n \vee 2)} + \frac{56L}{\theta r \sqrt{n}} \sqrt{k_n} s_n \sqrt{\ln(p_n \vee 2)}. \end{aligned}$$

□

Corollary 5.3. *Assume that $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma_{jn}^2)$, condition $C_\epsilon(w)$ holds for some $\epsilon, \theta > 0$ and for every $w \subseteq \{1, \ldots, p_n\}$ such that $|w| \leq k_n$, $k_n \ln(p_n \vee 2) = o(n)$ and $k_n \ln(p_n \vee 2) = o(a_n)$. Then we have*

$$\mathbb{P}(\min_{w \in \mathcal{M}: s^* \subset w} GIC(w) \leq GIC(s^*)) \to 0.$$

Proof. We take: $r_n = C_n \sqrt{\frac{k_n \ln(p_n \vee 2)}{n}}$, where $C_n = \sqrt[4]{\frac{n}{k_n \ln(p_n \vee 2)}} \min\{1, \sqrt[4]{\frac{a_n}{n}}\}$. We observe that $C_n \to +\infty$, $r_n \leq \sqrt[4]{\frac{k_n \ln(p_n \vee 2)}{n}} \to 0$ and

$$C_n \frac{k_n \ln(p_n \vee 2)}{a_n} \leq \left(\frac{k_n \ln(p_n \vee 2)}{a_n} \right)^{\frac{3}{4}} \to 0.$$

In view of Theorem 5.2 we have for sufficiently large n such that $r_n < \epsilon$ holds:

$$\begin{aligned} \mathbb{P}(\min_{w \in \mathcal{M}: s^* \subset w} GIC(w) \leq GIC(s^*)) &\leq \frac{14L\sqrt{k_n} s_n \sqrt{\ln(p_n \vee 2)} r_n \sqrt{n}}{a_n} + \frac{56L\sqrt{k_n} s_n \sqrt{\ln(p_n \vee 2)}}{\sqrt{n}\theta r_n} \\ &= \frac{14LC_n k_n s_n \ln(p_n \vee 2)}{a_n} + \frac{56L s_n}{\theta C_n} \to 0. \end{aligned}$$

□

The most restrictive condition of Corollary 5.3 is $k_n \ln(p_n \vee 2) = o(a_n)$. We note that in the case when $p_n \geq n$ and $k_n = d$, EBIC penalty defined above corresponds to the borderline of this condition.

Theorem 5.4. *Assume that $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma_{jn}^2)$, condition $C_\epsilon(s^*)$ holds for some $\epsilon, \theta > 0$ and $8a_n|s^*| < \theta n \min\{\epsilon^2, \beta_{min}^{*2}\}$. Then we have:*

$$\mathbb{P}(\min_{w \in \mathcal{M}: w \subset s^*} GIC(w) \leq GIC(s^*)) \leq \frac{32Ls_n\sqrt{|s^*|}}{\theta\sqrt{n}\min\{\epsilon, \beta_{min}^*\}}.$$

Proof. Suppose that for some $w \subset s^*$ we have $GIC(w) \leq GIC(s^*)$. This inequality is equivalent to:

$$nR_n(\hat{\boldsymbol{\beta}}(s^*)) - nR_n(\hat{\boldsymbol{\beta}}(w)) \geq a_n(|w| - |s^*|).$$

In view of inequalities $R_n(\hat{\boldsymbol{\beta}}(s^*)) \leq R_n(\boldsymbol{\beta}^*)$ and $a_n(|w| - |s^*|) \geq -a_n|s^*|$ we obtain:

$$nR_n(\boldsymbol{\beta}^*) - nR_n(\hat{\boldsymbol{\beta}}(w)) \geq -a_n|s^*|.$$

Let:

$$\mathbf{v} = u\hat{\boldsymbol{\beta}}(w) + (1-u)\boldsymbol{\beta}^*$$

for some $u \in [0, 1]$ specified later. From convexity of ρ we have:

$$nR_n(\boldsymbol{\beta}^*) - nR_n(\mathbf{v}) \geq nu(R_n(\boldsymbol{\beta}^*) - R_n(\hat{\boldsymbol{\beta}}(w))) \geq -ua_n|s^*| \geq -a_n|s^*|. \tag{5.10}$$

We have two cases:

1) $\beta_{min}^* > \epsilon$.

First we observe that exists some $h_0 \in (0, 1)$ such that

$$a_n|s^*| \leq \frac{\theta}{2} \left(\frac{h_0}{h_0 + 1} \right)^2 \epsilon^2 n, \tag{5.11}$$

what follows from our assumption. Let $h \in [h_0, 1)$, $r = h\epsilon$, $u = r/(r + ||\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^*||_2)$ and

$$\mathbf{v} = u\hat{\boldsymbol{\beta}}(w) + (1-u)\boldsymbol{\beta}^*. \tag{5.12}$$

Note that $||\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^*||_2 \geq ||\boldsymbol{\beta}^*_{s^* \setminus w}||_2 \geq \beta^*_{min}$. Then, as function $d(x) = x/(x+c)$ is increasing and bounded from above by 1 for $x, c > 0$, we obtain:

$$r = h\epsilon \geq ||\mathbf{v} - \boldsymbol{\beta}^*||_2 = \frac{h\epsilon||\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^*||_2}{h\epsilon + ||\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^*||_2} \geq \frac{h\epsilon\beta^*_{min}}{h\epsilon + \beta^*_{min}} \geq \frac{h\epsilon^2}{(h+1)\epsilon} = \frac{h}{h+1}\epsilon. \tag{5.13}$$

Hence, in view of $C_\epsilon(s^*)$ condition we have:

$$R(\mathbf{v}) - R(\boldsymbol{\beta}^*) \geq \theta \left(\frac{h}{h+1}\right)^2 \epsilon^2.$$

Using (5.10)-(5.12) and above inequality yields:

$$S_2(r) \geq R_n(\boldsymbol{\beta}^*) - R(\boldsymbol{\beta}^*) - (R_n(\mathbf{v}) - R(\mathbf{v})) \geq \theta \left(\frac{h}{h+1}\right)^2 \epsilon^2 - \frac{a_n}{n}|s^*| \geq \frac{\theta}{2}\left(\frac{h}{h+1}\right)^2 \epsilon^2.$$

Thus, in view of Lemma 5.1, we obtain:

$$\mathbb{P}(\min_{w \in \mathcal{M}: w \subset s^*} GIC(w) \leq GIC(s^*)) \leq \mathbb{P}\left(S_2(r) \geq \frac{\theta}{2}\left(\frac{h}{h+1}\right)^2 \epsilon^2\right) \leq \frac{8L\sqrt{|s^*|}s_n(h+1)^2}{\sqrt{n}\theta h\epsilon}.$$

Taking $h \to 1^-$ leads to inequality:

$$\mathbb{P}(\min_{w \in \mathcal{M}: w \subset s^*} GIC(w) \leq GIC(s^*)) \leq \frac{32L\sqrt{|s^*|}s_n}{\sqrt{n}\theta\epsilon}. \tag{5.14}$$

2) $\beta^*_{min} \leq \epsilon$.

In this case we take $u = \beta^*_{min}/(\beta^*_{min} + ||\hat{\boldsymbol{\beta}}(w) - \boldsymbol{\beta}^*||_2)$ and define $\mathbf{v}$ as in (5.12). Analogously as in (5.13), we have:

$$\frac{\beta^*_{min}}{2} \leq ||\mathbf{v} - \boldsymbol{\beta}^*||_2 \leq \beta^*_{min}.$$

Hence, in view of $C_\epsilon(s^*)$ condition we have:

$$R(\mathbf{v}) - R(\boldsymbol{\beta}^*) \geq \theta\frac{\beta^{*2}_{min}}{4}.$$

Using (5.10) and above inequality yields:

$$S_2(\beta^*_{min}) \geq R_n(\boldsymbol{\beta}^*) - R(\boldsymbol{\beta}^*) - (R_n(\mathbf{v}) - R(\mathbf{v})) \geq \theta\frac{\beta^{*2}_{min}}{4} - \frac{a_n}{n}|s^*| \geq \frac{\theta}{8}\beta^{*2}_{min}.$$

Thus, in view of Lemma 5.1, we obtain:

$$\mathbb{P}(\min_{w \in \mathcal{M}: w \subset s^*} GIC(w) \leq GIC(s^*)) \leq \mathbb{P}\left(S_2(\beta^*_{min}) \geq \frac{\theta}{8}\beta^{*2}_{min}\right) \leq \frac{32L\sqrt{|s^*|}s_n}{\sqrt{n}\theta\beta^*_{min}}. \tag{5.15}$$

By combining (5.14) and (5.15) the theorem follows. □

Corollary 5.5. *Assume that loss $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma^2_{jn})$, condition $C_\epsilon(s^*)$ holds for some $\epsilon, \theta > 0$ and $a_n|s^*| = o(n\min\{1, \beta^*_{min}\}^2)$, then*

$$\mathbb{P}(\min_{w \in \mathcal{M}: w \subset s^*} GIC(w) \leq GIC(s^*)) \to 0.$$

Proof. First we observe that if $a_n|s^*| = o(n\min\{1, \beta^*_{min}\}^2)$ and $|s^*| = o(n\min\{1, \beta^*_{min}\}^2)$, then in view of Theorem 5.4 we have

$$\mathbb{P}(\min_{w\in\mathcal{M}: w\subset s^*} GIC(w) \leq GIC(s^*)) \to 0.$$

Condition $|s^*| = o(n\min\{1, \beta^*_{min}\}^2)$ can be omitted as $a_n \to +\infty$ and is implied by $a_n|s^*| = o(n\min\{1, \beta^*_{min}\}^2)$. □

5.2. Selection consistency of SS procedure

SS (Screening and Selection) procedure for the model without intercept is the following:

1. Take some $\lambda > 0$.
2. Find $\hat{\boldsymbol{\beta}}_{L,-0} = \underset{\mathbf{b}\in\mathbb{R}^{p_n}}{\arg\min} R_n((0, \mathbf{b}^T)^T) + \lambda||\mathbf{b}||_1$.
3. Find $\hat{s}_L = \operatorname{supp}\hat{\boldsymbol{\beta}}_{L,-0} = \{j_1, \ldots, j_k\}$ such that $|\hat{\beta}_{L,-0,j_1}| \geq \ldots \geq |\hat{\beta}_{L,-0,j_k}| > 0$ and $j_1, \ldots, j_k \in \{1, \ldots, p_n\}$.
4. Define $\mathcal{M}_{SS} = \{\emptyset, \{j_1\}, \{j_1, j_2\}, \ldots, \{j_1, j_2, \ldots, j_k\}\}$.
5. Find $\hat{s} = \underset{w\in\mathcal{M}_{SS}}{\arg\min} GIC(w)$.

SS procedure is a modification of SS procedure in Pokarowski et al. (2018).

In the model with intercept we modify SS procedure as follows:

1. Take some $\lambda > 0$.
2. Find $\hat{\boldsymbol{\beta}}_L = \underset{\mathbf{b}\in\mathbb{R}^{p_n+1}}{\arg\min} R_n(\mathbf{b}) + \lambda||\tilde{\mathbf{b}}||_1$.
3. Find $\hat{s}_L = \operatorname{supp}\hat{\boldsymbol{\beta}}_L \setminus \{0\} = \{j_1, \ldots, j_k\}$ such that $|\hat{\beta}_{L,j_1}| \geq \ldots \geq |\hat{\beta}_{L,j_k}| > 0$ and $j_1, \ldots, j_k \in \{1, \ldots, p_n\}$.
4. $\mathcal{M}_{SS} = \{\emptyset, \{j_1\}, \{j_1, j_2\}, \ldots, \{j_1, j_2, \ldots, j_k\}\}$.
5. Find $\hat{s} = \underset{w\in\mathcal{M}_{SS}}{\arg\min} GIC(w \cup \{0\})$.

Corollaries 5.6, 5.7 and Remark 5.8 describe the situations when SS procedure works.

Corollary 5.6. *(model without intercept) Assume that $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma^2_{jn})$ and $\boldsymbol{\beta}^*$ exists and is unique. If $k_n \in \mathbb{N}_+$ is some sequence, margin condition (MC) is satisfied for some $\vartheta, \delta, \varepsilon > 0$, condition $C_\epsilon(w)$ holds for some $\epsilon, \theta > 0$ and for every $w \subseteq \{1, \ldots, p_n\}$ such that $|w| \leq k_n$, $\mathcal{M}_{SS}$ is hierarchical family constructed in the step 4 of SS procedure and the following conditions are fulfilled:*

- $\mathbb{P}(\forall w \in \mathcal{M}_{SS} : |w| \leq k_n) \to 1$,
- $|s^*| \leq k_n$,
- $\liminf_n \kappa_{\boldsymbol{H}}(\varepsilon) > 0$ *for some $\varepsilon > 0$, where $\boldsymbol{H}$ is non-negative definite matrix and $\kappa_{\boldsymbol{H}}(\varepsilon)$ is defined in (4.15),*

- $\log(p_n) = o(n\lambda^2)$,
- $k_n\lambda = o(\min\{\beta^*_{min}, 1\})$,
- $k_n \log p_n = o(n)$,
- $k_n \log p_n = o(a_n)$,
- $a_n k_n = o(n \min\{\beta^*_{min}, 1\}^2)$,

then for SS procedure for the model without intercept we have

$$\mathbb{P}(\hat{s} = s^*) \to 1.$$

Proof. In view of Corollary 4.16 from separation property (4.31) we obtain $\mathbb{P}(s^* \in \mathcal{M}_{SS}) \to 1$. Let:

$$A_1 = \{\min_{w \in \mathcal{M}_{SS}: w \supset s^*, |w| \le k_n} GIC(w) \le GIC(s^*)\},$$
$$A_2 = \{\min_{w \in \mathcal{M}_{SS}: w \supset s^*, |w| > k_n} GIC(w) \le GIC(s^*)\},$$
$$B = \{\forall w \in \mathcal{M}_{SS} : |w| \le k_n\}.$$

Then we have again from union inequality:

$$\begin{aligned}\mathbb{P}(\min_{w \in \mathcal{M}_{SS}: w \supset s^*} GIC(w) \le GIC(s^*)) &= \mathbb{P}(A_1 \cup A_2) \\ &= \mathbb{P}((A_1 \cap B) \cup (A_1 \cap B^c) \cup (A_2 \cap B) \cup (A_2 \cap B^c)) \\ &\le \mathbb{P}(A_1 \cap B) + \mathbb{P}(A_1 \cap B^c) + \mathbb{P}(A_2 \cap B) + \mathbb{P}(A_2 \cap B^c).\end{aligned}$$

Firstly we observe that $\mathbb{P}(A_2 \cap B) = 0$. In view of Corollary 5.3 and monotonicity of probability we obtain:

$$\mathbb{P}(A_1 \cap B) \le \mathbb{P}(A_1) \to 0.$$

Similarly, we obtain:

$$\mathbb{P}(A_1 \cap B^c) \le \mathbb{P}(B^c) \to 0,$$
$$\mathbb{P}(A_2 \cap B^c) \le \mathbb{P}(B^c) \to 0.$$

Hence:

$$\mathbb{P}(\min_{w \in \mathcal{M}_{SS}: w \supset s^*} GIC(w) \le GIC(s^*)) \to 0. \tag{5.16}$$

In the analogous way, using $|s^*| \le k_n$ and Corollary 5.5 yields:

$$\mathbb{P}(\min_{w \in \mathcal{M}_{SS}: w \subset s^*} GIC(w) \le GIC(s^*)) \to 0. \tag{5.17}$$

Now, observe that in view of definition of $\hat{s}$ and union inequality:

$$\begin{aligned}\mathbb{P}(\hat{s} = s^*) &= \mathbb{P}(\min_{w \in \mathcal{M}_{SS}: w \ne s^*} GIC(w) > GIC(s^*)) \\ &\ge 1 - \mathbb{P}(\min_{w \in \mathcal{M}_{SS}: w \subset s^*} GIC(w) \le GIC(s^*)) - \mathbb{P}(\min_{w \in \mathcal{M}_{SS}: w \supset s^*} GIC(w) \le GIC(s^*)).\end{aligned}$$

Thus $\mathbb{P}(\hat{s} = s^*) \to 1$ in view of above inequality, (5.16) and (5.17). □

Corollary 5.7. *(logistic model with intercept) Assume that ρ is logistic loss, $X_{ij} \sim Subg(\sigma_{jn}^2)$ and $\boldsymbol{\beta}^*$ exists and is unique. If $k_n \in \mathbb{N}$ is some sequence, condition $C_\epsilon(w)$ holds for some $\epsilon, \theta > 0$ and for every $w \subseteq \{1, \ldots, p_n\}$ such that $|w_0| \leq k_n$, where $w_0 = w \cup \{0\}$, $\mathcal{M}_{SS}$ is hierarchical family constructed in the step 4 of SS procedure and the following conditions are fulfilled:*

- $m \leq \kappa \leq M$ *(κ defined in (4.13)) for some $m, M > 0$,*
- $|s_0^*| \leq k_n$,
- $\mathbb{P}(\forall w \in \mathcal{M}_{SS}\colon\ |w_0| \leq k_n) \to 1$,
- $k_n^2 \log p_n = o(n)$,
- $\lambda^2 k_n^2 \log(np_n) = o(1)$,
- $\log p_n = o(n\lambda^2)$,
- $k_n \lambda^2 = o(\beta_{min}^{*2})$,
- $k_n \log p_n = o(a_n)$,
- $a_n k_n = o(n \min\{\beta_{min}^*, 1\}^2)$.

then for SS procedure for the logistic model with intercept we have:

$$\mathbb{P}(\hat{s} = s^*) \to 1.$$

Proof. Proof of this Corollary is analogous to the proof of Corollary 5.6 and it follows from Corollaries 4.10, 5.3 and 5.5. We note that condition $k_n^2 \log p_n = o(n)$ from Corollary 5.6 implies $k_n \log p_n = o(n)$ (as $k_n \geq 1$), what is one of assumptions in Corollary 5.3. □

Remark 5.8. *If $p_n = O(e^{cn^\gamma})$ for some $c > 0$, $\gamma \in (0, 1/2)$, $\xi \in (0, 0.5 - \gamma)$, $u \in (0, 0.5 - \gamma - \xi)$, $k_n = O(n^\xi)$, $\lambda = C_n\sqrt{\log(p_n)/n}$, $C_n = O(n^u)$, $C_n \to +\infty$, $n^{-\frac{u}{2}} = O(\beta_{min}^*)$, $a_n = dn^{\frac{1}{2}-u}$, then assumptions about asymptotic behaviour of parameters in Corollary 5.7 are satisfied.*

Remark 5.9. *We note that in order to apply Corollary 5.7 to two-step procedure based on Lasso it is required that $|s_0^*| \leq k_n$ and that the support of Lasso estimator with probability tending to 1 contains no more than k_n elements. Some results bounding $|\operatorname{supp} \hat{\boldsymbol{\beta}}_L|$ are available for deterministic $\boldsymbol{X}$ (see Huang et al. (2008)) and for random $\boldsymbol{X}$ (see Theorems A.27, A.30), but they are too weak to be useful in our context, therefore we use stronger assumptions. The other possibility to prove consistency of two-step procedure is to modify it in the first step by using thresholded Lasso (see Zhou (2010)) corresponding to k_n' largest Lasso coefficients where $k_n' \in \mathbb{N}$ is such that $k_n = o(k_n')$.*

Chapter 6

Numerical experiments

6.1. Logistic loss - calculation of $\boldsymbol{\beta}^*$

Note that finding explicit form of projections $\boldsymbol{\beta}^*$ is rarely possible in continuous case. Nevertheless, our objective here will be to show that for a given distribution of r.v. $\mathbf{X}$ and function q computation of projection $\boldsymbol{\beta}^*$ is numerically feasible for logistic loss. Reasoning presented here can be easily generalized to other loss functions.

6.1.1. General assumptions

To compute $\boldsymbol{\beta}^*$ in general case, we define: $F(\mathbf{b}) = \mathbb{E}q_L(\mathbf{b}^T\mathbf{X})\mathbf{X} - \mathbb{E}Y\mathbf{X}$. In view of normal equations (2.10): $F(\boldsymbol{\beta}^*) = \mathbf{0}_p$. Now we compute matrix of the first derivatives of F: $J_F(\mathbf{b}) = \mathbb{E}q'_L(\mathbf{b}^T\mathbf{X})\mathbf{X}\mathbf{X}^T$. If $\mathbf{X}$ satisfies LND and $\mathbb{E}||\tilde{\mathbf{X}}||_2^2 < \infty$, then $J_F(\mathbf{b})$ is well defined positive definite matrix, as we have for every $\mathbf{b} \in \mathbb{R}^{p+1}, \gamma \in \mathbb{R}^{p+1} \setminus \{\mathbf{0}_{p+1}\}$: $\gamma^T J_F(\mathbf{b})\gamma = \mathbb{E}q'_L(\mathbf{b}^T\mathbf{X})||\gamma^T\mathbf{X}||^2 > 0$. The iteration of Newton–Raphson method is thus given by:

$$\boldsymbol{\beta}^*_{n+1} = \boldsymbol{\beta}^*_n - J_F(\boldsymbol{\beta}^*_n)^{-1}F(\boldsymbol{\beta}^*_n).$$

6.1.2. Generalized semiparametric model - linear regressions condition

If $\mathbf{X}$ satisfies linear regressions condition and

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q(\mathbf{B}^T\mathbf{X}), \tag{6.1}$$

then we know that $\tilde{\boldsymbol{\beta}}^* = \tilde{\mathbf{B}}\boldsymbol{\eta}$, hence $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\eta}^T\tilde{\mathbf{B}}^T)^T$ and we need to estimate only β_0^* and $\boldsymbol{\eta}$, what is much easier task than under general assumptions. In this case we define

$$F(x, \mathbf{y}) = \begin{bmatrix} \mathbb{E}q_L(x + \mathbf{y}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) - \mathbb{E}q(\mathbf{B}^T\mathbf{X}) \\ \mathbb{E}q_L(x + \mathbf{y}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}})\tilde{\mathbf{B}}^T\tilde{\mathbf{X}} - \mathbb{E}q(\mathbf{B}^T\mathbf{X})\tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \end{bmatrix}.$$

Matrix of the first derivatives of F has the form:

$$J_F(x, \mathbf{y}) = \begin{bmatrix} \mathbb{E}q'_L(x + \mathbf{y}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) & \mathbb{E}q'_L(x + \mathbf{y}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}})\tilde{\mathbf{X}}^T\tilde{\mathbf{B}} \\ \mathbb{E}q'_L(x + \mathbf{y}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}})\tilde{\mathbf{B}}^T\tilde{\mathbf{X}} & \mathbb{E}q'_L(x + \mathbf{y}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}})\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\tilde{\mathbf{B}} \end{bmatrix}.$$

Similarly as in general case, $J_F(x,y) > 0$ and we can use Newton–Raphson iterations:

$$\begin{bmatrix} \beta^*_{0,n+1} \\ \boldsymbol{\eta}_{n+1} \end{bmatrix} = \begin{bmatrix} \beta^*_{0,n} \\ \boldsymbol{\eta}_n \end{bmatrix} - J_F(\beta^*_{0,n}, \boldsymbol{\eta}_n)^{-1} F(\beta^*_{0,n}, \boldsymbol{\eta}_n).$$

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

$$\mathbb{E}q(\beta_0 + \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) = \mathbb{E}q_L(\beta_0^* + \boldsymbol{\eta}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}}) \approx q_L(\beta_0^*)$$

and using Remark 3.16 (we use normality of $\tilde{\mathbf{X}}$ here):

$$\boldsymbol{\eta} = \frac{\mathbb{E}Dq(\beta_0 + \tilde{\mathbf{B}}^T\tilde{\mathbf{X}})}{\mathbb{E}q_L'(\beta_0^* + \boldsymbol{\eta}^T\tilde{\mathbf{B}}^T\tilde{\mathbf{X}})} \approx \frac{\mathbb{E}Dq(\beta_0 + \tilde{\mathbf{B}}\tilde{\mathbf{X}})}{q_L'(\beta_0^*)}.$$

Hence we can take:

$$\beta^*_{0,0} = q_L^{-1}(\mathbb{E}q(\beta_0 + \tilde{\mathbf{B}}^T\tilde{\mathbf{X}})),$$

$$\boldsymbol{\eta}_0 = \frac{\mathbb{E}Dq(\beta_0 + \tilde{\mathbf{B}}^T\tilde{\mathbf{X}})}{q_L'(\beta^*_{0,0})} = \frac{\mathbb{E}q'(\beta_0 + \tilde{\mathbf{B}}^T\tilde{\mathbf{X}})}{\mathbb{E}q(\beta_0 + \tilde{\mathbf{B}}^T\tilde{\mathbf{X}})(1 - \mathbb{E}q(\beta_0 + \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}))},$$

as $q_L'(\beta^*_{0,0}) = q_L(\beta^*_{0,0})(1 - q_L(\beta^*_{0,0})) = \mathbb{E}q(\beta_0 + \tilde{\mathbf{B}}^T\tilde{\mathbf{X}})(1 - \mathbb{E}q(\beta_0 + \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}))$.

We note that the above procedure is similar to procedure for semiparametric model - we replace $\mathbf{B}$ by $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ by η. Moreover, if $\tilde{\mathbf{X}}$ does not follow normal distribution with zero mean, we choose starting point equal zero instead of the one given above.

6.2. Simulation I - calculation of $\boldsymbol{\beta}^*$ in semiparametric model

We assume that $\mathbf{X} = (1, \tilde{\mathbf{X}}^T)^T$, where $\tilde{\mathbf{X}} \sim \mathcal{N}_p(\mathbf{0}_p, I_p)$ and $p = 15$. Conditional distribution $Y|\mathbf{X}$ is given by semiparametric model:

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q(\boldsymbol{\beta}_s^T\mathbf{x}_s).$$

The following coefficients have been considered (they are the same as in the numerical experiments in Mielniczuk and Teisseyre (2016)):

(M1) $s = \{10\}$, $\boldsymbol{\beta}_s = 0.2$,
(M2) $s = \{2,4,5\}$, $\boldsymbol{\beta}_s = (1,1,1)^T$,
(M3) $s = \{1,2\}$, $\boldsymbol{\beta}_s = (0.5, 0.7)^T$,
(M4) $s = \{1,2\}$, $\boldsymbol{\beta}_s = (0.3, 0.5)^T$,
(M5) $s = \{1,\ldots,8\}$, $\boldsymbol{\beta}_s = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)^T$.

Let $\Phi(\cdot)$ denote distribution function of standard normal distribution 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:

$$q_1(s) = \Phi(s),$$

$$q_2(s) = \begin{cases} \Phi(s) & \text{for } \Phi(s) \in (0.1, 0.8) \\ 0.1 & \text{for } \Phi(s) \leq 0.1 \\ 0.8 & \text{for } \Phi(s) \geq 0.8, \end{cases}$$

$$q_3(s) = \begin{cases} \Phi(s) & \text{for } \Phi(s) \in (0.2, 0.7) \\ 0.2 & \text{for } \Phi(s) \leq 0.2 \\ 0.7 & \text{for } \Phi(s) \geq 0.7, \end{cases}$$

$$q_4(s) = \begin{cases} \Phi(s) & \text{for } |s| > 1 \\ 0.5 + 0.5\cos[4\pi s]\Phi(s) & \text{for } |s| \leq 1, \end{cases}$$

$$q_5(s) = F_{Cauchy(0,1)}(s),$$

$$q_6(s) = F_{Cauchy(0,2)}(s).$$

In Mielniczuk and Teisseyre (2016) values of η were calculated using Monte Carlo experiments and are given in Table 1 there. In Tables 6.1-6.2 values of β_0^* and η 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. Comparing these results with simulated values in Mielniczuk and Teisseyre (2016), we observe that for all functions except q_4 (non-monotonic case) the results of both calculations are very close, what suggest that the above Newton-Raphson procedure performs well for monotone q. Note that Monte-Carlo calculation of $\hat{\eta}$ performed in Mielniczuk and Teisseyre (2016) was based on $n = 10^6$ observations drawn from distribution of $(\mathbf{X}, Y)$ whereas here we use a single run of the iterative procedure for its evaluation. We also note that values of η (and $\hat{\eta}$) for q_1 are greater but close to $\sqrt{8/\pi}$, what is consistent with (2.42).

Table 6.1: Values of η 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 6.2: Values of β_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 6.3: Simulated values of $\hat{\eta}$ for considered models reproduced from Mielniczuk and Teisseyre (2016) (first 5 rows) together with MSEs between simulated and numerically 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

6.3. Simulation II - calculation of $\boldsymbol{\beta}^*$ in additive binary model

We consider $\mathbf{X} = (X_1, X_2)^T \sim \mathcal{N}_2(\mathbf{0}_2, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} r^2 & \rho r \\ \rho r & 1 \end{bmatrix},$$

$r \geq 1$ and $\rho \in (-1, 1)$. We assume that:

$$\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) = \alpha q_L(x_1) + (1 - \alpha) q_L(x_2)$$

for $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$. This means that $\boldsymbol{\beta}_1 = (0, 1, 0)^T, \boldsymbol{\beta}_2 = (0, 0, 1)^T$. We fit the model with logistic loss without intercept. In view of Theorem 3.8:

$$\boldsymbol{\beta}^* = \tilde{\mathbf{B}}\boldsymbol{\eta} = \eta_1 \tilde{\boldsymbol{\beta}}_1 + \eta_2 \tilde{\boldsymbol{\beta}}_2 = (\eta_1, \eta_2)^T.$$

It follows from Lemma A.51 part 4 that function $h(\sigma) = \mathbb{E} q_L'(\sigma Z)$ is decreasing for $Z \sim \mathcal{N}(0, 1)$ as its derivative $\mathbb{E} Z q_L''(\sigma Z)$ is negative. This means that $\mathbb{E} q_L'(X_1) \leq \mathbb{E} q_L'(X_2)$, because $\operatorname{Var} X_1 \geq \operatorname{Var} X_2$. Thus Corollary 3.24 gives us:

$$\eta_1 \leq \alpha,$$
$$\eta_2 \leq (1 - \alpha) \frac{\mathbb{E} q_L'(X_2)}{\mathbb{E} q_L'(X_1)}.$$

When $r = 1$, then X_1 and X_2 have the same distribution and thus $\eta_2 \leq 1 - \alpha$. This yields $\eta_1 + \eta_2 \leq 1$, when $r = 1$. For $r = 2$ lower panels of Figure 6.1 suggest that $\eta_1 + \eta_2 \leq 1$, but it is still open problem to prove it for $r \neq 1$. Moreover, in the case $\rho = 0.9$ the difference $1 - \eta_1 - \eta_2$ is smaller than in the case $\rho = 0$.

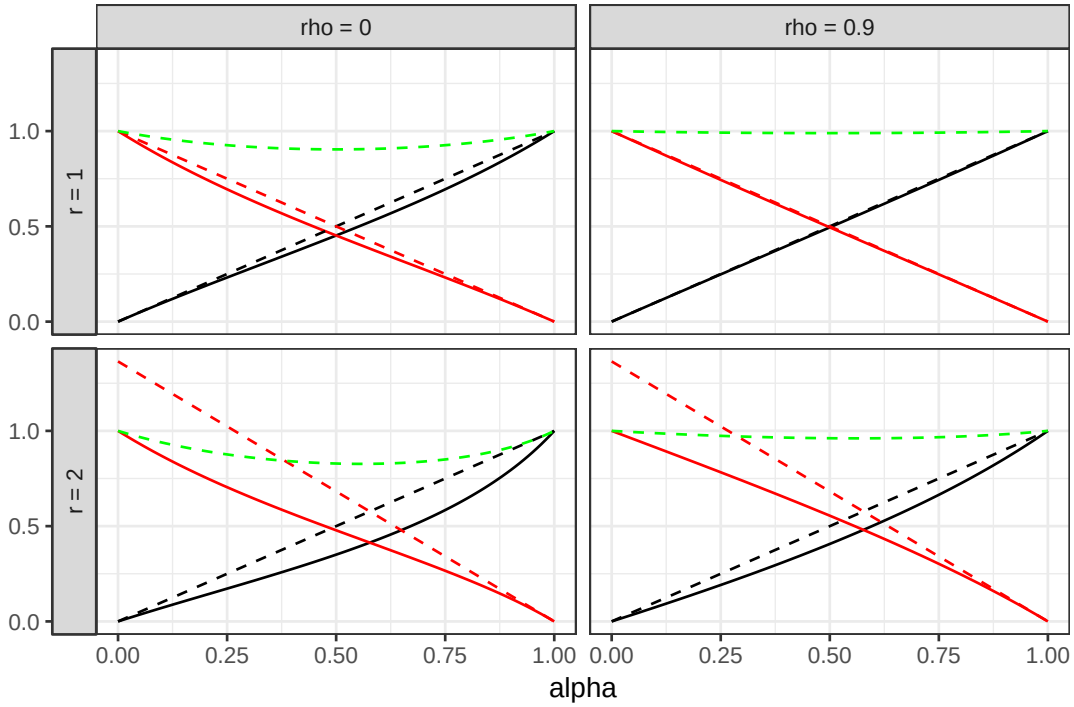


Figure 6.1: Values of η_1 and η_2 against α. Black solid line shows the values of η_1 versus α, black dashed line represents upper bound for $\eta_1 = \alpha$ (see text). Red lines correspond to η_2 and upper bound for η_2. Dashed green line represents sum of η_1 and η_2.

Figure 6.1 contains graph computed values of η_1 and η_2 as a functions of α for $r = 1, 2$ and $\rho = 0, 0.9$. η_1 and η_2 were computed with Newton–Raphson method (see Section 6.1.1) and expected values approximated by Gauss–Hermite quadrature with 50 nodes (using R package `fastGHQuad`).

6.4. Selection procedures

In performed simulations we have implemented modifications of SS procedure introduced in Section 5.2, as the original procedure is defined for a single λ only. In practice it is generally easier to consider some sequence $\lambda_1 > \ldots > \lambda_m > 0$ instead of λ in the first step, because we do not know how to choose the best λ (see e.g. Remark 4.11). When we consider the sequence $\lambda_1, \ldots, \lambda_m$, we can construct for corresponding families $\mathcal{M}_1, \ldots, \mathcal{M}_m$ having the similar form to $\mathcal{M}$ in the step 4 of SS procedure. Hence we arrive here at the following SSnet procedure, which is the modification of SOSnet procedure in Pokarowski et al. (2018):

1. Choose some $\lambda_1 > \ldots > \lambda_m > 0$.
2. Find $\hat{\boldsymbol{\beta}}_L^{(i)} = \underset{\mathbf{b}\in\mathbb{R}^{p_n+1}}{\arg\min}\, R_n(\mathbf{b}) + \lambda_i||\tilde{\mathbf{b}}||_1$ for $i = 1, \ldots, m$.
3. Find $\hat{s}_L^{(i)} = \operatorname{supp} \hat{\tilde{\boldsymbol{\beta}}}_L^{(i)} = \{j_1^{(i)}, \ldots, j_{k_i}^{(i)}\}$ where $j_1^{(i)}, \ldots, j_{k_i}^{(i)}$ are such that $|\hat{\beta}^{(i)}_{L,j_1^{(i)}}| \geq \ldots \geq |\hat{\beta}^{(i)}_{L,j_k^{(i)}}| > 0$ for $i = 1, \ldots, m$.
4. Define $\mathcal{M}_i = \{\{j_1^{(i)}\}, \{j_1^{(i)}, j_2^{(i)}\}, \ldots, \{j_1^{(i)}, j_2^{(i)}, \ldots, j_{k_i}^{(i)}\}\}$ for $i = 1, \ldots, m$.
5. Define $\mathcal{M} = \{\emptyset\} \cup \bigcup_{i=1}^{m} \mathcal{M}_i$.
6. Find $\hat{s} = \underset{w\in\mathcal{M}}{\arg\min}\, GIC(w \cup \{0\})$, where $GIC(w \cup \{0\}) = \min_{\mathbf{b}\in\mathbb{R}^{p_n+1}:\operatorname{supp}\tilde{\mathbf{b}}\subseteq w} nR_n(\mathbf{b}) + a_n(|w| + 1)$.

Instead of constructing families $\mathcal{M}_i$ for each λ_i in SSnet procedure, we can choose λ_i by cross-validation using "one-standard error" rule (see Friedman et al. (2010)) and then proceed as in SS procedure. This gives the following SSCV procedure:

1. Take some $\lambda_1 > \ldots > \lambda_m > 0$.
2. For each $i = 1, \ldots, m$ compute for Lasso model K-fold cross-validation error $E_{CV}(\lambda_i)$ and a standard deviation of cross-validation error $SD_{CV}(\lambda_i)$.
3. Find $\lambda_{min} = \arg\min_{\lambda_1,\ldots,\lambda_m} E_{CV}(\lambda_i)$.
4. Choose $\lambda = \max\{\lambda_i\colon\ E_{CV}(\lambda_i) \leq E_{CV}(\lambda_{min}) + SD_{CV}(\lambda_{min}), i = 1, \ldots, m\}$.
5. Find $\hat{\boldsymbol{\beta}}_L = \underset{\mathbf{b}\in\mathbb{R}^{p_n+1}}{\arg\min}\, R_n(\mathbf{b}) + \lambda||\tilde{\mathbf{b}}||_1$.
6. Find $\hat{s}_L = \operatorname{supp} \hat{\tilde{\boldsymbol{\beta}}}_L = \{j_1, \ldots, j_k\}$ such that $|\hat{\beta}_{L,j_1}| \geq \ldots \geq |\hat{\beta}_{L,j_k}| > 0$.
7. Define $\mathcal{M} = \{\emptyset, \{j_1\}, \{j_1, j_2\}, \ldots, \{j_1, j_2, \ldots, j_k\}\}$.
8. Find $\hat{s} = \underset{w\in\mathcal{M}}{\arg\min}\, GIC(w \cup \{0\})$, where $GIC(w \cup \{0\}) = \min_{\mathbf{b}\in\mathbb{R}^{p_n+1}:\operatorname{supp}\tilde{\mathbf{b}}\subseteq w} nR_n(\mathbf{b}) + a_n(|w| + 1)$.

The last procedure considered in this dissertation has been introduced in Fan and Tang (2013) and contains also a step for choosing λ. However, it is different from previous procedures, as λ is chosen by minimizing function similar to GIC, but computed for Lasso estimator instead of MLE. Then $\hat{s}$ is equal to support of optimal Lasso estimator. We took $a_n = \log(\log n) \cdot \log(p_n + 1)$ as in Fan and Tang (2013). We will call this procedure LFT:

1. Take some $\lambda_1 > \ldots > \lambda_m > 0$.
2. Find $\hat{\boldsymbol{\beta}}_L^{(i)} = \underset{\mathbf{b} \in \mathbb{R}^{p_n+1}}{\arg\min}\, R_n(\mathbf{b}) + \lambda_i ||\tilde{\mathbf{b}}||_1$ for $i = 1, \ldots, m$.
3. Find $\hat{s}_L^{(i)} = \{j \in \{0, 1, \ldots, p_n\} \colon \hat{\beta}_{L,j}^{(i)} \neq 0\}$ for $i = 1, \ldots, m$.
4. Find $i_0 = \arg\min_{i=1,\ldots,m} nR_n(\hat{\boldsymbol{\beta}}_L^{(i)}) + a_n |\hat{s}_L^{(i)}|$, where $a_n = \log(\log n) \cdot \log(p_n + 1)$.
5. Find $\hat{s} = \hat{s}_L^{(i_0)} \setminus \{0\}$.

We note that family $\mathcal{M}$ is defined for LFT procedure (in order to compute performance measures) as:

$$\mathcal{M} = \{\hat{s}_L^{(i)} \colon \ i = 1, \ldots, m\}.$$

We list below versions of the above procedures along with R packages, which were used to choose sequence $\lambda_1, \ldots, \lambda_m$ and computation of Lasso estimator. The following packages were chosen based on selection performance after initial tests for each loss and procedure:

- SSnet with logistic or quadratic loss: `ncvreg`,
- SSCV or LFT with logistic or quadratic loss: `glmnet`,
- SSnet, SSCV or LFT with Huber loss: `hqreg`.

We list below functions which were used to optimize R_n in GIC minimization step for each loss:

- logistic loss: `glm.fit` (package `stats`),
- quadratic loss: `.lm.fit` (package `stats`),
- Huber loss: `rlm` (package `rlm`).

We did not perform simulations for probit loss and quantile loss due to time constraints and lack of well implemented R packages to compute Lasso estimator and MLE estimator for these loss functions.

Before applying each procedure, each column of matrix $\tilde{\mathbb{X}}$ was standardized, because $\hat{\boldsymbol{\beta}}_L$ depends on scaling of predictors. We set length of λ_i sequence to $m = 20$. Moreover, in all of the procedures we considered only λ_i for which $|\hat{s}_L^{(i)}| \leq n$. It is due to the fact that when $|\hat{s}_L^{(i)}| > n$ Lasso solutions are not unique (see discussion in Section A.3). For Huber loss we set parameter $\delta = 1/10$ (see Yi and Huang (2017)). Number of folds in SSCV was set to $K = 10$.

Each simulation consisted of L repetitions, during which samples $\mathbb{X}_k = (\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_n^{(k)})^T$ and $\mathbf{Y}_k = (Y_1^{(k)}, \ldots, Y_n^{(k)})^T$ were generated for $k = 1, \ldots, L$. For k-th sample $(\mathbb{X}_k, \mathbf{Y}_k)$ we have computed $\hat{s}_k$ - estimator of set of active predictors obtained by a given procedure,

$$\hat{\boldsymbol{\beta}}(\hat{s}_k) = (\hat{\beta}_0(\hat{s}_k), \hat{\tilde{\boldsymbol{\beta}}}(\hat{s}_k)^T)^T = \underset{\mathbf{b} \in \mathbb{R}^{p_n+1}: \mathbf{b}_{(\hat{s}_k \cup \{0\})^c}=0}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho(\mathbf{b}^T \mathbf{X}_i^{(k)}, Y_i^{(k)})$$

is MLE estimator for k-th sample on set $\hat{s}_k$.

$\mathcal{M}^{(k)}$ is the family $\mathcal{M}$ obtained by a given procedure for k-th sample.

In our numerical experiments we have computed the following measures of selection performance:

- $ANGLE = \frac{1}{L} \sum_{k=1}^{L} \arccos |\cos \angle(\tilde{\boldsymbol{\beta}}, \hat{\tilde{\boldsymbol{\beta}}}(\hat{s}_k))|$, where $\cos \angle(\tilde{\boldsymbol{\beta}}, \hat{\tilde{\boldsymbol{\beta}}}(\hat{s}_k)) = \frac{\sum_{j=1}^{p_n} \beta_j \hat{\beta}_j(\hat{s}_k)}{||\tilde{\boldsymbol{\beta}}||_2 ||\hat{\tilde{\boldsymbol{\beta}}}(\hat{s}_k)||_2}$ and we let $\cos \angle(\tilde{\boldsymbol{\beta}}, \hat{\tilde{\boldsymbol{\beta}}}(\hat{s}_k)) = 0$, if $||\tilde{\boldsymbol{\beta}}||_2 ||\hat{\tilde{\boldsymbol{\beta}}}(\hat{s}_k)||_2 = 0$,
- $P_{inc} = \frac{1}{L} \sum_{k=1}^{L} I(s^* \in \mathcal{M}^{(k)})$,
- $P_{equal} = \frac{1}{L} \sum_{k=1}^{L} I(\hat{s}_k = s^*)$,
- $P_{supset} = \frac{1}{L} \sum_{k=1}^{L} I(\hat{s}_k \supseteq s^*)$.

In our simulations we additionally computed time of 1st stage of each procedure, which includes finding Lasso estimators and building family $\mathcal{M}$ (in case of SSnet and SSCV) and time of 2nd stage which includes GIC minimization.

6.5. Simulation III - selection

6.5.1. Experimental setup - model M1

We generated n observations $(\mathbf{X}_i, Y_i) \in \mathbb{R}^{p+1} \times \{0, 1\}$ for $i = 1, \ldots, n$ such that:

$$X_{i1} = Z_{i1}, X_{i2} = Z_{i2}, X_{ij} = Z_{i,j-7} \text{ for } j = 10, \ldots, p,$$

$$X_{i3} = X_{i1}^2, X_{i4} = X_{i2}^2, X_{i5} = X_{i1}X_{i2}, X_{i6} = X_{i1}^2 X_{i2}, X_{i7} = X_{i1}X_{i2}^2, X_{i8} = X_{i1}^3, X_{i9} = X_{i2}^3,$$

where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = [\rho^{|i-j|}]_{i,j=1,\ldots,p}$ and $\rho \in (-1, 1)$. We consider response function $q(x) = q_L(x^3)$ for $x \in \mathbb{R}$, $s = \{1, 2\}$ and $\boldsymbol{\beta}_s = (1, 1)^T$. This means that:

$$\begin{aligned}\mathbb{P}(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i) &= q(\boldsymbol{\beta}_s^T \mathbf{x}_{i,s}) = q(x_{i1} + x_{i2}) = q_L((x_{i1} + x_{i2})^3) \\ &= q_L(x_{i1}^3 + x_{i2}^3 + 3x_{i1}^2 x_{i2} + 3x_{i1} x_{i2}^2) = q_L(3x_{i6} + 3x_{i7} + x_{i8} + x_{i9}).\end{aligned}$$

We observe that the above binary model is well specified with respect to family of logistic models. Hence $s_{log}^* = \{6, 7, 8, 9\}$ and $\boldsymbol{\beta}_{log, s_{log}^*}^* = (3, 3, 1, 1)^T$ are respectively set of active predictors and non-zero coefficients of projection onto family of logistic models.

We took the following parameters in the simulation: $n = 500, p = 150, \rho \in \{-0.9+0.15\cdot k\colon\ k = 0, 1, \ldots, 12\}$ and $L = 500$ - number of generated data sets for each combination of parameters. We considered procedures SSnet, SSCV and LFT using logistic, quadratic and Huber loss functions. For procedures SSnet and SSCV we used GIC penalties with:

- $a_n = \log n$ (BIC),
- $a_n = \log n + 2 \log p_n$ (EBIC1).

6.5.2. Experimental setup - model M2

We generated n observations $(\mathbf{X}_i, Y_i) \in \mathbb{R}^{p+1} \times \{0, 1\}$ for $i = 1, \ldots, n$ such that $\tilde{\mathbf{X}}_i = (X_{i1}, \ldots, X_{ip})^T \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$, $\mathbf{\Sigma} = [\rho^{|i-j|}]_{i,j=1,\ldots,p}$ and $\rho \in (-1, 1)$. We took response function $q(x) = q_L(x^3)$ for $x \in \mathbb{R}$, $s = \{1, 2\}$ and $\boldsymbol{\beta}_s = (1, 1)^T$. This means that:

$$\mathbb{P}(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i) = q(\boldsymbol{\beta}_s^T \mathbf{x}_{i,s}) = q(x_{i1} + x_{i2}) = q_L((x_{i1} + x_{i2})^3)$$

This model in comparison to the one presented in Section 6.5.1 does not contain monomials of X_{i1} and X_{i2} of degree higher than 1 in its set of predictors. We observe that this binary model is missspecified with respect to family of logistic models, because $q(x_{i1} + x_{i2}) \not\equiv q_L(\boldsymbol{\beta}^T \mathbf{x}_i)$ for any $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. However, in this case linear regressions condition LRC is satisfied for $\tilde{\mathbf{X}}$, as it follows normal distribution. Hence in view of Remark 2.24 we have $s^*_{log} = \{1, 2\}$ and $\boldsymbol{\beta}^*_{log, s^*_{log}} = \eta(1, 1)^T$ for some $\eta > 0$.

We took the following parameters in the simulation: $n = 500, p = 150, \rho \in \{-0.9+0.15\cdot k\colon\ k = 0, 1, \ldots, 12\}$ and $L = 500$ - number of generated data sets for each combination of parameters. We considered procedures SSnet, SSCV and LFT using logistic, quadratic and Huber loss functions. For procedures SSnet and SSCV we used GIC penalties with:

- $a_n = \log n$ (BIC),
- $a_n = \log n + 2 \log p_n$ (EBIC1).

6.5.3. Results for models M1 and M2

We observe that values of P_{inc} for SSCV and SSnet are close to 1 for low correlations in M2 model for every tested loss (see Figure 6.2). In M1 model P_{inc} attains highest values for SSnet procedure and logistic loss for low correlations - this is due to the fact that in most cases corresponding family $\mathcal{M}$ is the largest among the families created in considered procedures. P_{inc} is close to 0 in model M1 for quadratic and Huber loss, what affects other measures and could be caused by large correlations in the data, as we have $\mathrm{Cor}(X_{i1}, X_{i8}) = 3/\sqrt{15} \approx 0.77$. It is seen that in model M1 inclusion probability P_{inc} is much lower than in model M2 (except for negative correlations). It it also seen that P_{inc} for SSCV is larger than for LFT and LFT fails with respect to P_{inc} in M1.

In the model M1 P_{equal} attains highest values for SSnet with BIC penalty, then for SSCV with EBIC1 penalty (see Figure 6.3). In the model M2 P_{equal} attains values close to 1 for SSnet and SSCV with EBIC1 penalty and was much larger than P_{equal} for the corresponding versions using BIC penalty - moreover choice of loss was significant only for larger correlations. These results confirm theoretical results of Theorem 2.6. We observe also that although in the model M2 remaining procedures do not select s^* with high probability, they select $w \supset s^*$ with high probability, what is indicated by values of P_{supset} (see 6.4). This analysis is confirmed by an analysis of $ANGLE$ measure (see 6.5), which attains values close to 0, when P_{supset} is close to 1. Low values of $ANGLE$ measure mean that estimated vector $\hat{\tilde{\boldsymbol{\beta}}}(\hat{s}_k))$ is approximately proportional to $\tilde{\boldsymbol{\beta}}$, what was the case for M2 model, where we had normal predictors satisfying linear regressions condition. Note that $\hat{\tilde{\boldsymbol{\beta}}}(\hat{s}_k))$ and $\tilde{\boldsymbol{\beta}}^*$ are not approximately collinear in M1 despite the fact that M1 is well specified. Also, for the best performing procedures, P_{equal} was much larger in M2 than in M1, despite the fact that the latter is correctly specified. However there are 4 active variables in M1 compared to 2 in M2.

In model M1 procedures with BIC penalty performed better than those with EBIC1 penalty, however the gain for P_{equal} was much smaller than the gain when using EBIC1 in M2. LFT procedure performed poorly in model M1 and reasonably well in model M2. The overall winner in both model is SSnet. SSCV performs only slightly worse than SSnet in M2 but it is significantly worse in M1.

Analysis of computing times of 1st and 2nd stage of each procedure shows that SSnet procedure creates large families $\mathcal{M}$, thus GIC minimization becomes computationally intensive. We also observe that 1st stage for SSCV takes more time than for SSnet, what is caused by multiple fitting of Lasso in cross-validation. However, SSCV is much quicker than SSnet in 2nd stage. We do not compare times between M1 and M2 models, because the results were calculated on 2 separate machines.

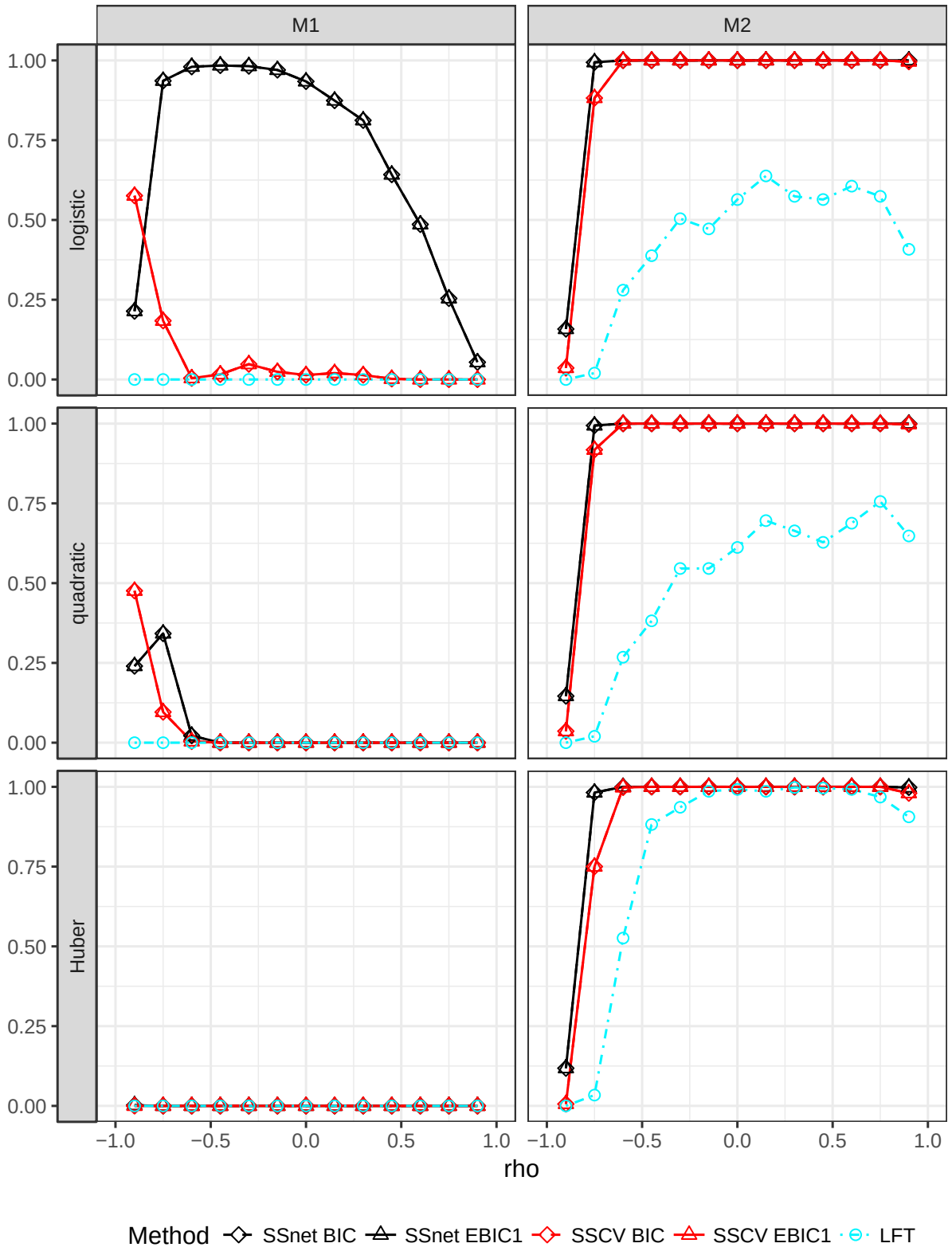


Figure 6.2: P_{inc} for models M1 and M2

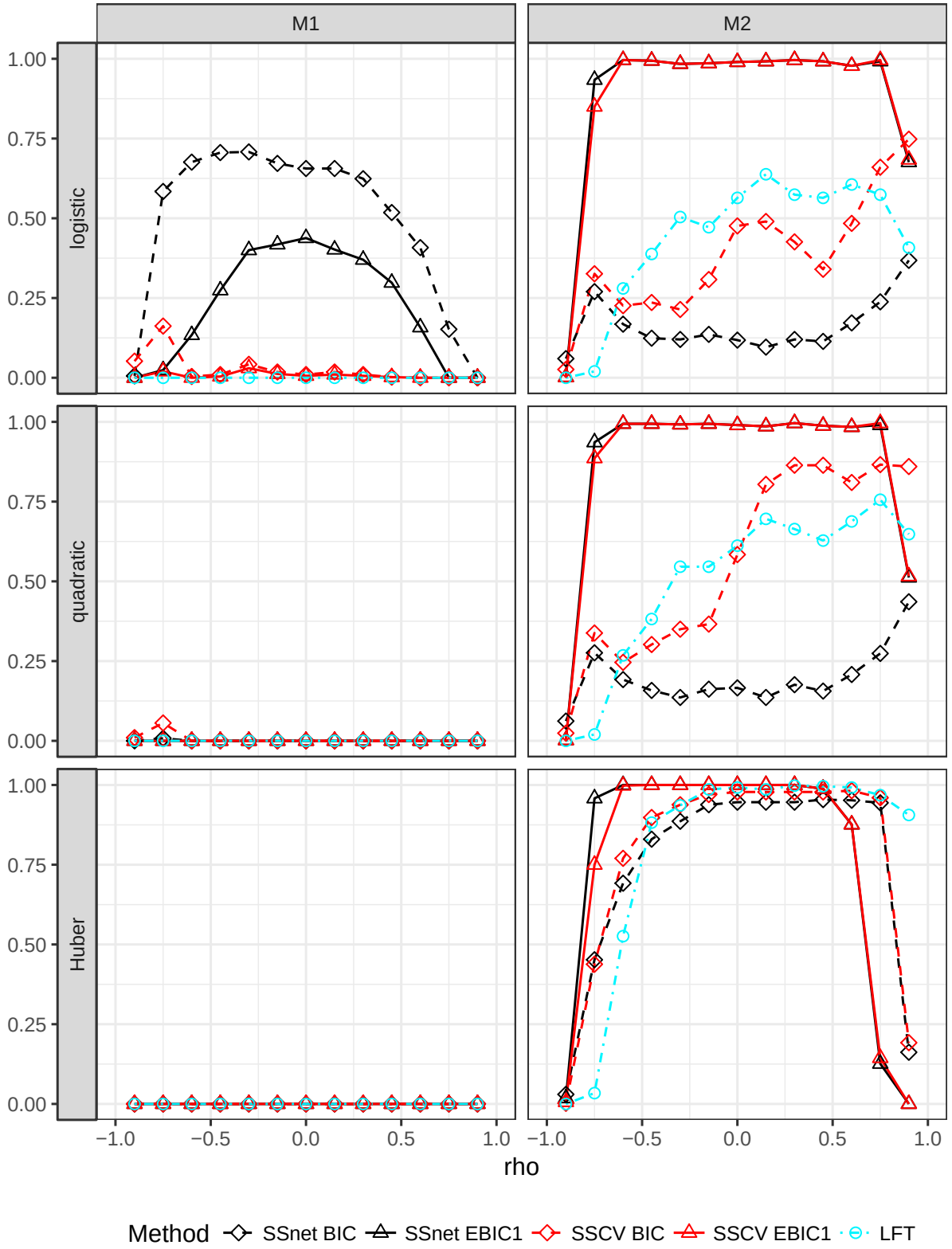


Figure 6.3: P_{equal} for models M1 and M2

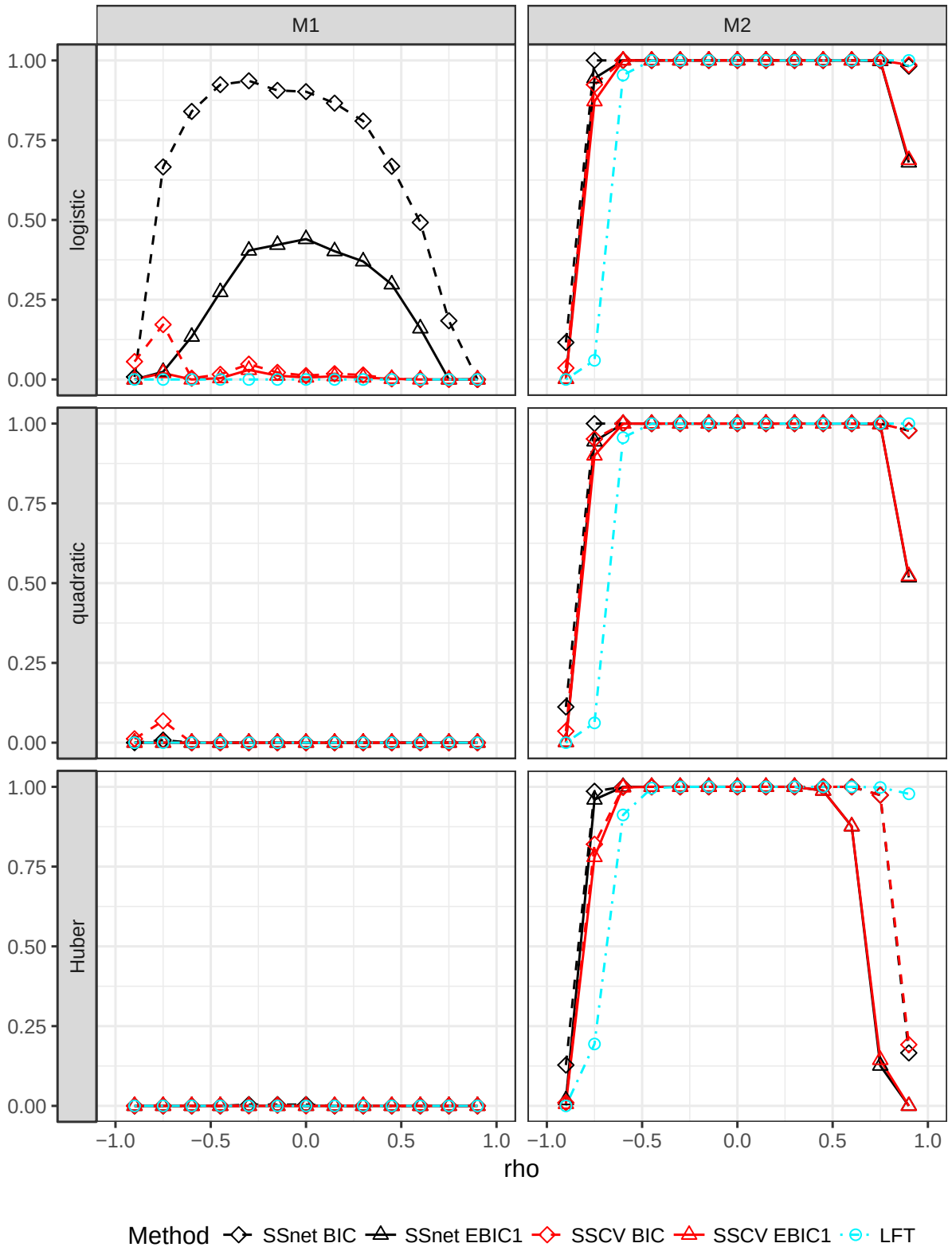


Figure 6.4: P_{supset} for models M1 and M2

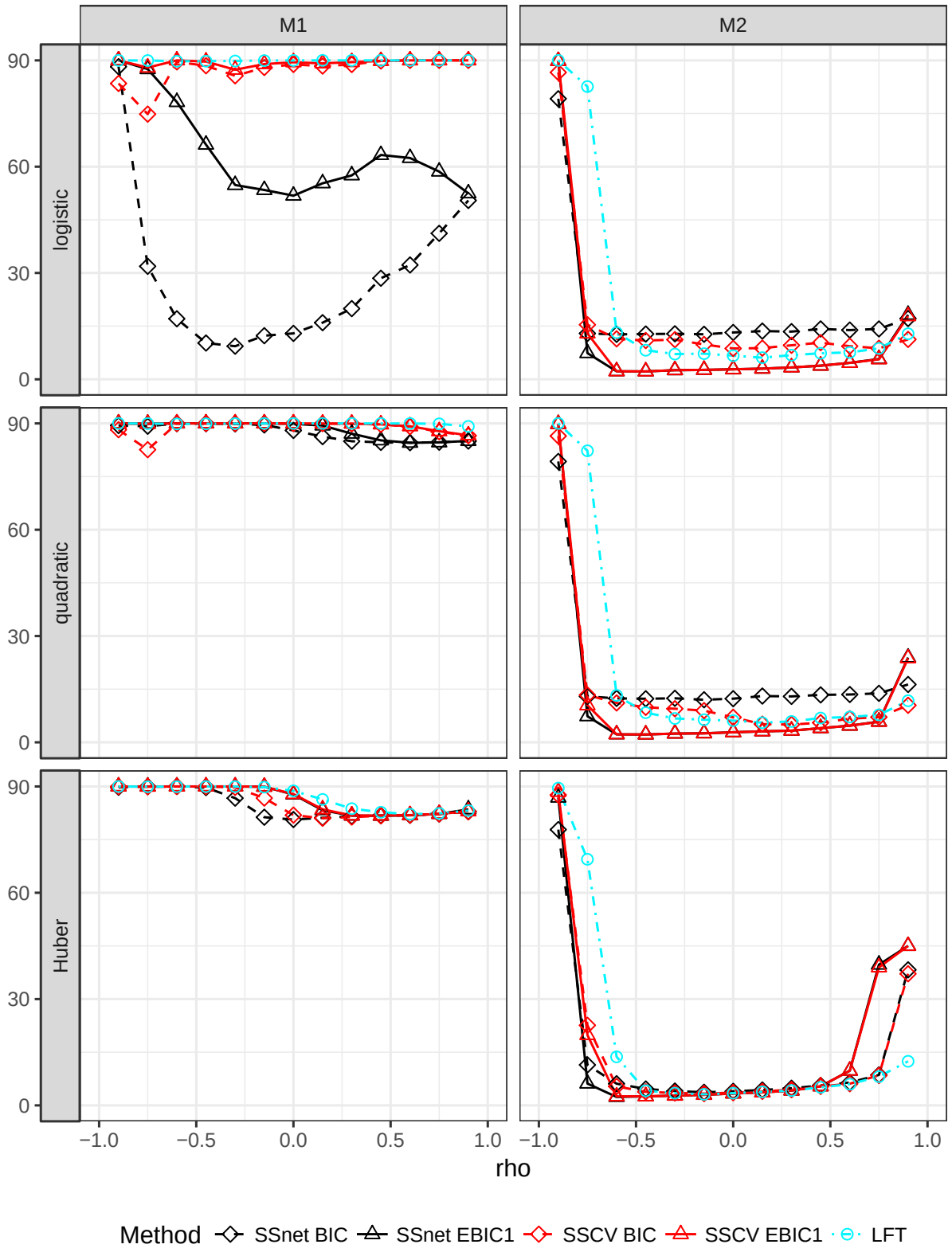


Figure 6.5: *ANGLE* for models M1 and M2

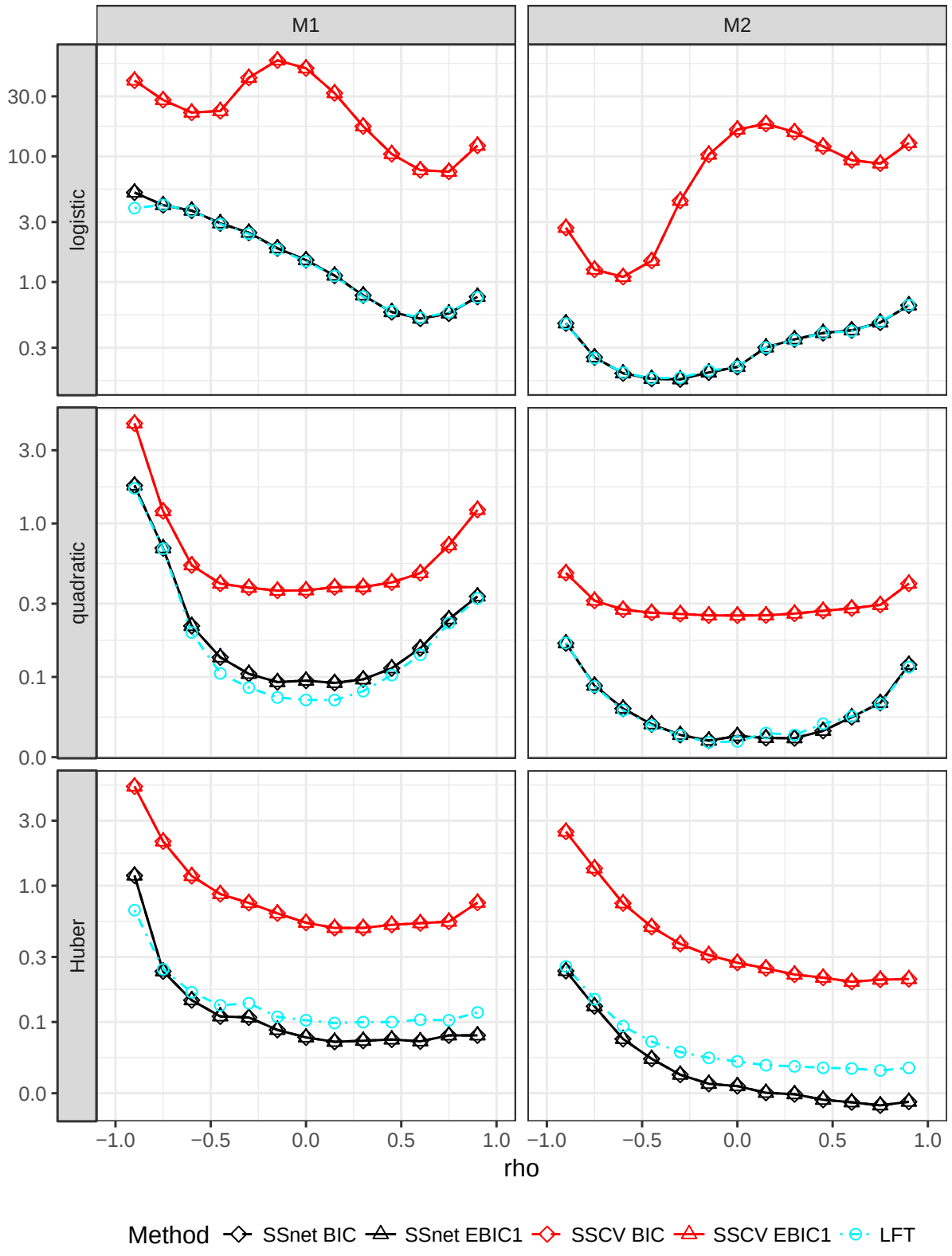


Figure 6.6: Time of selecting λ and building family $\mathcal{M}$ for models M1 and M2 (logarithmic scale y)

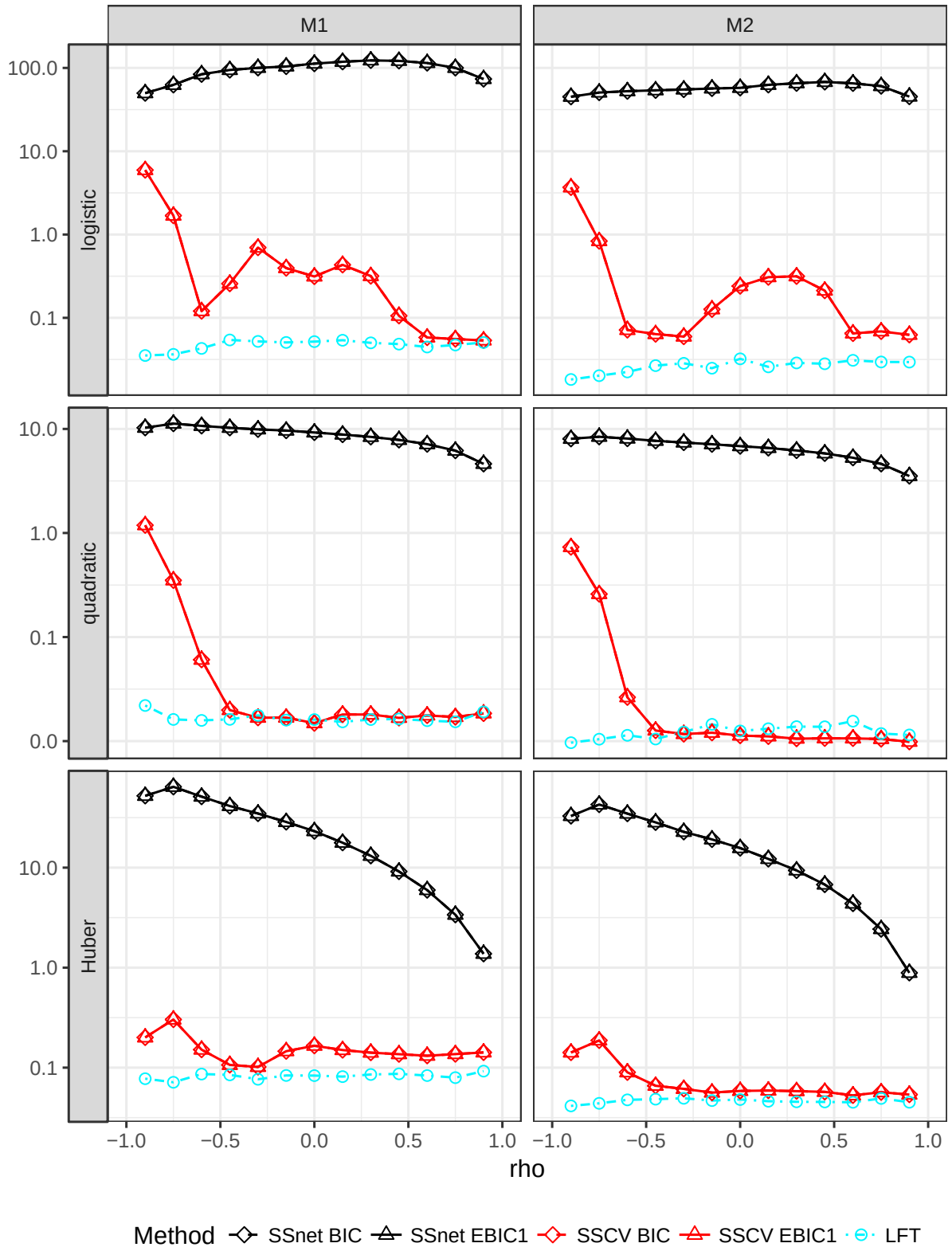


Figure 6.7: Time of GIC minimizing for models M1 and M2 (logarithmic scale y)

6.5.4. Experimental setup - model M2a

To check robustness of procedures in this chapter, we considered also modification of model M2, where $\mathbf{X}$ do not follow $\mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$ distribution with $\mathbf{\Sigma} = [\rho^{|i-j|}]_{i,j=1,\ldots,p}$ and $\rho \in (-1, 1)$. Below we define the modification for model M2a and vector $\mathbf{X}$.

Firstly, we observe that vector $(X_1, \ldots, X_p)^T$ having $\mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$ distribution and $\mathbf{\Sigma} = [\rho^{|i-j|}]_{i,j=1,\ldots,p}$ is AR(1) process constructed as follows:

$$X_1 = \varepsilon_1$$
$$X_j = \rho X_{j-1} + \sqrt{1-\rho^2}\varepsilon_j \text{ for } j = 2, \ldots, p,$$

where $\varepsilon_1, \ldots, \varepsilon_p \sim \mathcal{N}(0, 1)$ are independent. We replace here $\mathcal{N}(0, 1)$ distribution of ε_i by $0.9\mathcal{N}(0, 1) + 0.05\mathcal{N}(5, 1) + 0.05\mathcal{N}(-5, 1)$. Moreover, because we want $\mathbf{X}$ to be (approximately) AR(1) process, we reject first k X_i generated by the above procedure (warm start). This means that the final algorithm for generating vector $\mathbf{X}$ is the following:

1. Generate $\varepsilon_1, \ldots, \varepsilon_{p+k} \sim 0.9\mathcal{N}(0, 1) + 0.05\mathcal{N}(5, 1) + 0.05\mathcal{N}(-5, 1)$.
2. Let $U_1 = \varepsilon_1, U_j = \rho U_{j-1} + \sqrt{1-\rho^2}\varepsilon_j$ for $j = 2, \ldots, p+k$.
3. Let $X_j = U_{j+k}$ for $j = 1, \ldots, p$.

In model M2a we consider $k = 100$.

6.5.5. Results for the model M2a

From the results it can be seen that P_{inc} is close to 1 even for large correlations for SSnet and SSCV (see Figure 6.8). LFT procedure performs poorly compared to SSnet and SSCV, when P_{inc} is considered. Moreover, P_{equal} attains highest values for SSnet with EBIC1 penalty and SSCV with EBIC1 is only slightly worse. P_{equal} in almost all situations (except $|\rho| = 0.9$) is close to 1 for SSnet and SSCV with EBIC1 penalty. P_{equal} for SSnet and SSCV with BIC penalty is lower than for these procedures with EBIC1 penalty. Moreover, P_{equal} for SSCV with BIC penalty is higher than for SSnet with BIC penalty. P_{equal} attains similar values for LFT and SSCV with BIC penalty for low correlations. P_{subset} attains high values (close to 1, especially for low correlations) for every method, except LFT with $\rho = -0.9$. This means that supersets of s^* are selected with high probability. Similarly, $ANGLE$ measure is lower than 20° for every method, except LFT with $\rho = -0.9$. This means that $\hat{\tilde{\boldsymbol{\beta}}}(\hat{s}_k)$ is approximately proportional to $\tilde{\boldsymbol{\beta}}$ although linear regressions condition does not hold in this case. The results for SSnet and SSCV with EBIC1 are similar for logistic and quadratic loss. It is worth noting that SSnet and SSCV procedures with quadratic loss are much faster than their versions with logistic loss (see Figure 6.9).

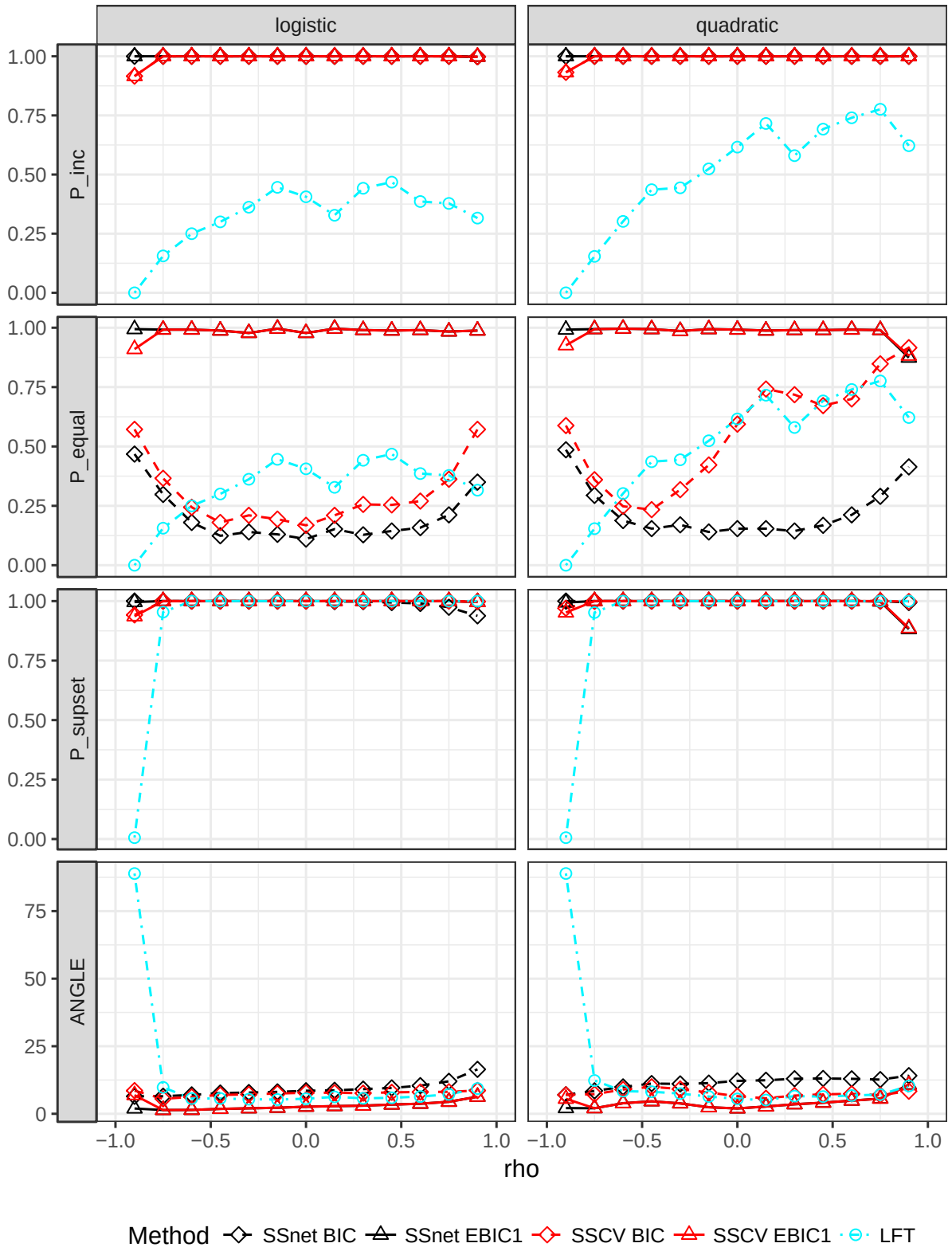


Figure 6.8: P_{equal}, P_{inc}, P_{supset} and $ANGLE$ for model M2a

Also, results of Simulation I suggest that fitting logistic model to a binary model with response function different from logistic may yield better results when the set of active predictors is sparse than for correctly specified model with larger number of potential predictors. Moreover, not much is lost in regard to probability of correct selection when linear model is fitted in place of logistic one in case of low correlations between predictors.

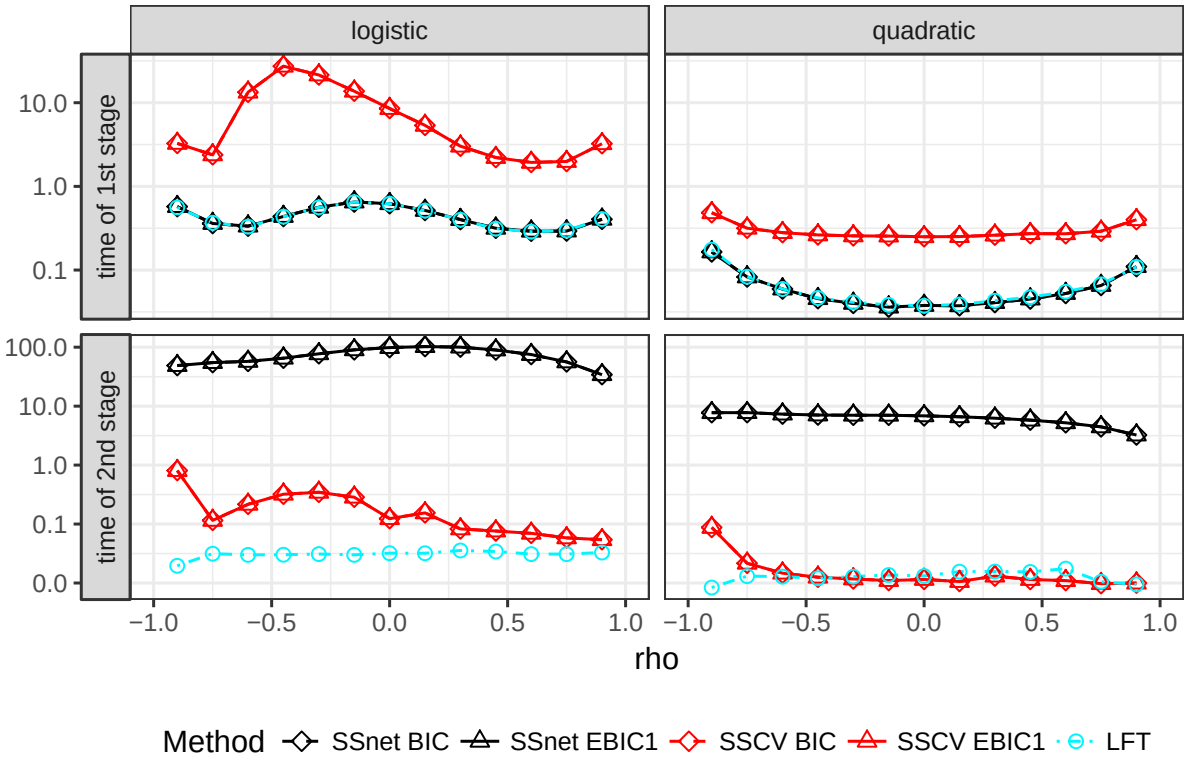


Figure 6.9: Times of 1st and 2nd stage of selection procedures for model M2a

6.6. Simulation IV - selection

6.6.1. Experimental setup - models MF1-MF4

We generated n observations $(\mathbf{X}_i, Y_i) \in \mathbb{R}^{p+1} \times \{0,1\}$ for $i = 1, \ldots, n$ such that $\tilde{\mathbf{X}}_i = (X_{i1}, \ldots, X_{ip})^T \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma})$, $\mathbf{\Sigma} = [\rho^{|i-j|}]_{i,j=1,\ldots,p}$ and $\rho \in (-1,1)$. We took response function q, $s \subseteq \{1, \ldots, p_n\}$ and $\boldsymbol{\beta}_s \in \mathbb{R}^{|s|}$ such that:

$$\mathbb{P}(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i) = q(\boldsymbol{\beta}_s^T \mathbf{x}_{i,s}).$$

Parameters n, p, s and $\boldsymbol{\beta}_s$, which we considered in the simulation are shown in the Table 6.4. Moreover, we took $\rho \in \{-0.9 + 0.15 \cdot k \colon \ k = 0, 1, \ldots, 12\}$ and $L = 500$ - number of generated data sets for each combination of parameters. We considered the following response functions:

- $q(x) = q_L(x)$,
- $q(x) = \Phi(x)$,
- $q(x) = F_{Cauchy}(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{atan} x$.

This setup for logistic loss, $q(x) = q_L(x)$, $\rho = 0$ and parameter values specified in Table 6.4 was considered in Fan and Tang (2013). Response functions given here were considered in Mielniczuk and Teisseyre (2016).

We considered procedures SSnet, SSCV and LFT using logistic loss function. For procedures SSnet and SSCV we used GIC penalties with:

- $a_n = \log n$ (BIC),
- $a_n = \log n + 2 \log p_n$ (EBIC1).

Model	n	p	s	$\boldsymbol{\beta}_s$
MF1	100	168	$\{1, 2, 5\}$	$(-3.5, 1.5, -2)^T$
MF2	180	692	$\{1, 2, 5, 6\}$	$(-3.5, 1.5, -2, 2)^T$
MF3	260	1993	$\{1, 2, 5, 6, 7\}$	$(-3.5, 1.5, -2, 2, -2)^T$
MF4	340	4680	$\{1, 2, 5, 6, 7, 8\}$	$(-3.5, 1.5, -2, 2, -2, 2)^T$

Table 6.4: Values of parameters in the Simulation 2

In this simulation we compared only values of P_{equal}, P_{inc} and P_{subset} due to limited space.

6.6.2. Results for models MF1-MF4

P_{inc} achieves highest values for negative correlations, moreover it increases with n for low correlations (see Figures 6.10, 6.13, 6.16). This affects P_{equal}, which achieves highest values for low correlations and increases with the sample size - the only exception was LFT procedure for which P_{equal} attained highest values for negative correlations (see Figures 6.11, 6.14, 6.17). P_{supset} attains significantly higher values than P_{equal} only for SSCV with BIC ponalty and LFT method (see Figures 6.12, 6.15, 6.18). From the results we observe that the model with response $q = \Phi$ was the easiest when logistic response was fitted. Moreover, the data with $q = F_{Cauchy}$ represented the most difficult case.

In this experiment it is seen that when P_{equal} is considered SSnet with EBIC1 is the overall winner for sufficiently high ρ ($\rho \geq -0.3$ for $q = q_L$, $\rho \geq -0.6$ for $q = \Phi$, $\rho \geq 0$ for $q = F_{Cauchy}$), and SSCV with EBIC1 performs only slightly worse. Moreover P_{equal} increases with n for low correlations. However, for large negative correlations LFT performs better in terms of P_{equal} than other procedures. Penalty change from EBIC1 to BIC results in very significant deterioration of performance measured by P_{inc} of both SSnet and SSCV. This means that the choice of penalty is crucial for performance of such selection procedures. Surprisingly, selection procedures performed better in overall for the probit i.e. misspecified binary model than for correctly specified logistic regression. We conclude that in the considered experiments SSnet with EBIC1 penalty works the best in most cases, however even for the winning procedure strong dependence of predictors makes the problem considerably harder.

It is clear from our experiments that choice of GIC penalty is crucial for its performance. Moreover, modification of SS procedure, which would perform satisfactorily for large correlations is still an open problem.

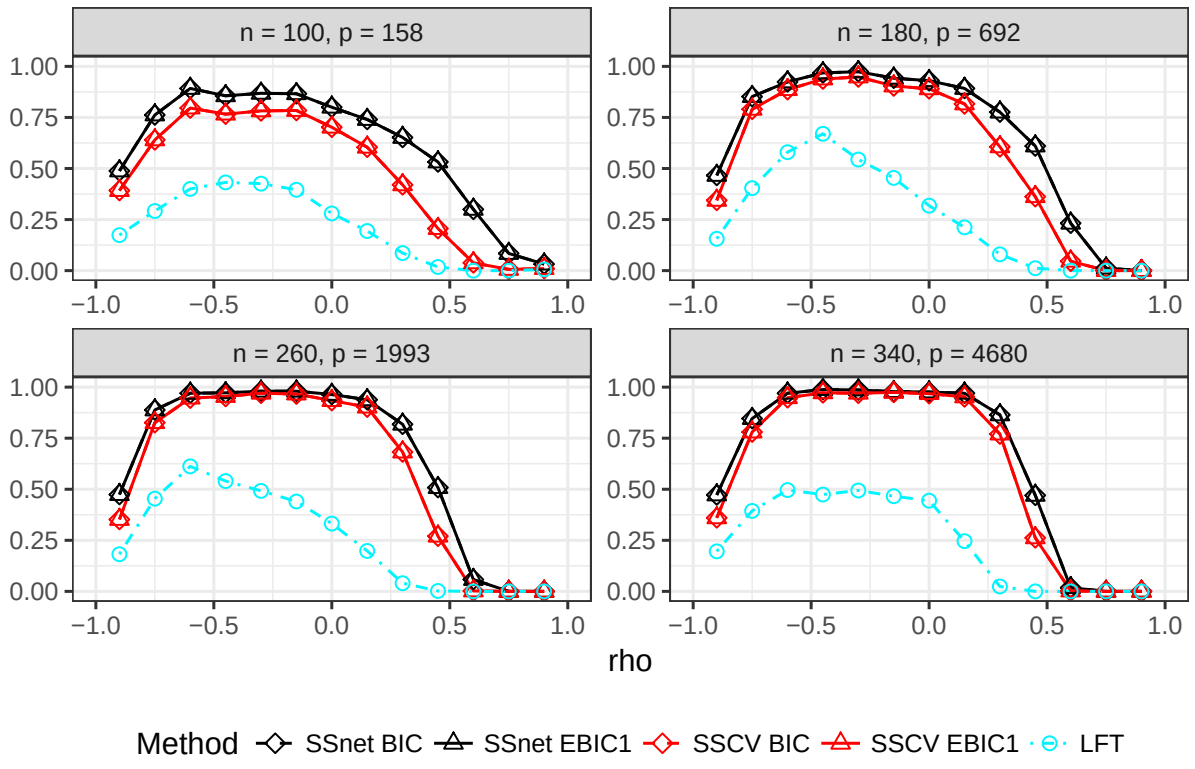


Figure 6.10: P_{inc} for models MF1-MF4 with $q = q_L$

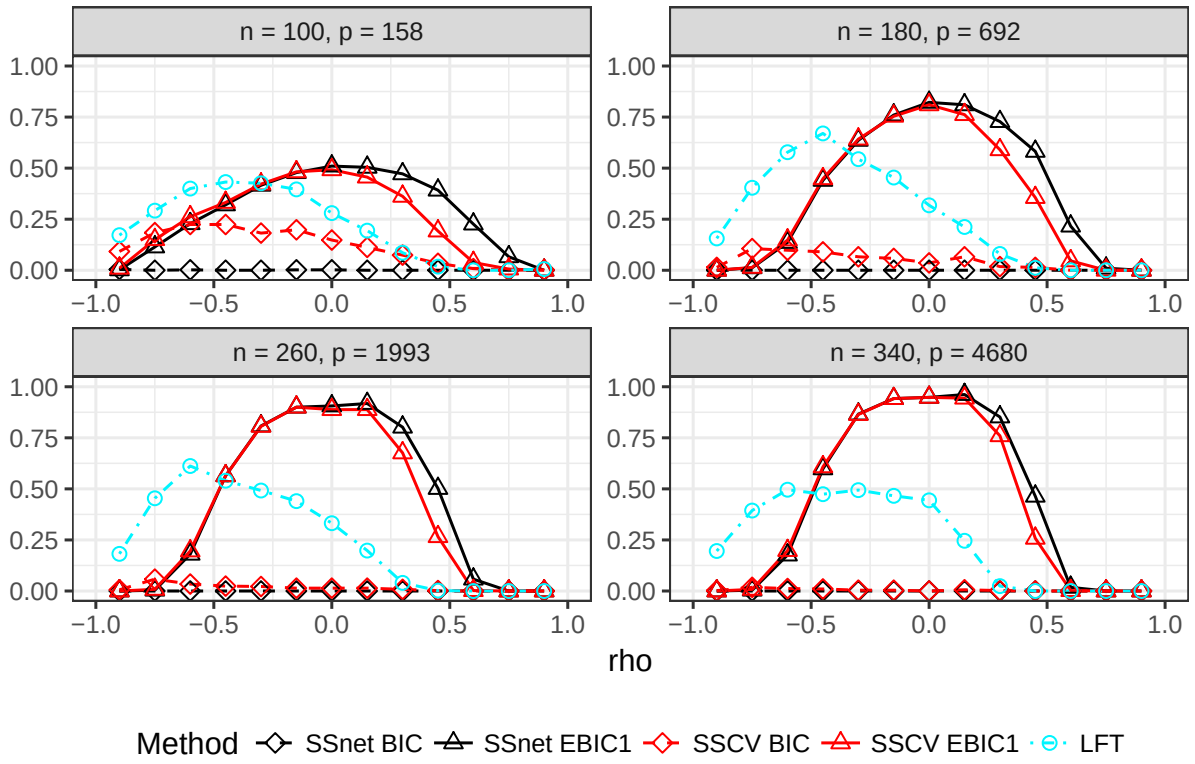


Figure 6.11: P_{equal} for models MF1-MF4 with $q = q_L$

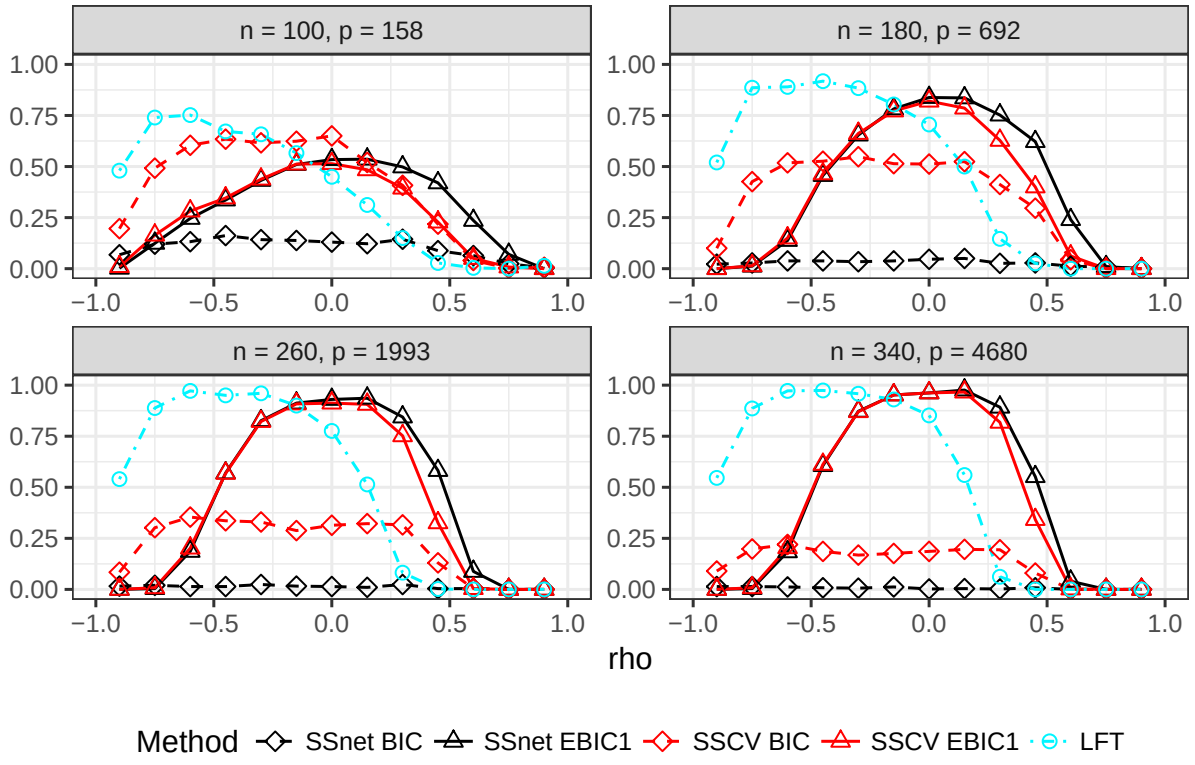


Figure 6.12: P_{supset} for models MF1-MF4 with $q = q_L$

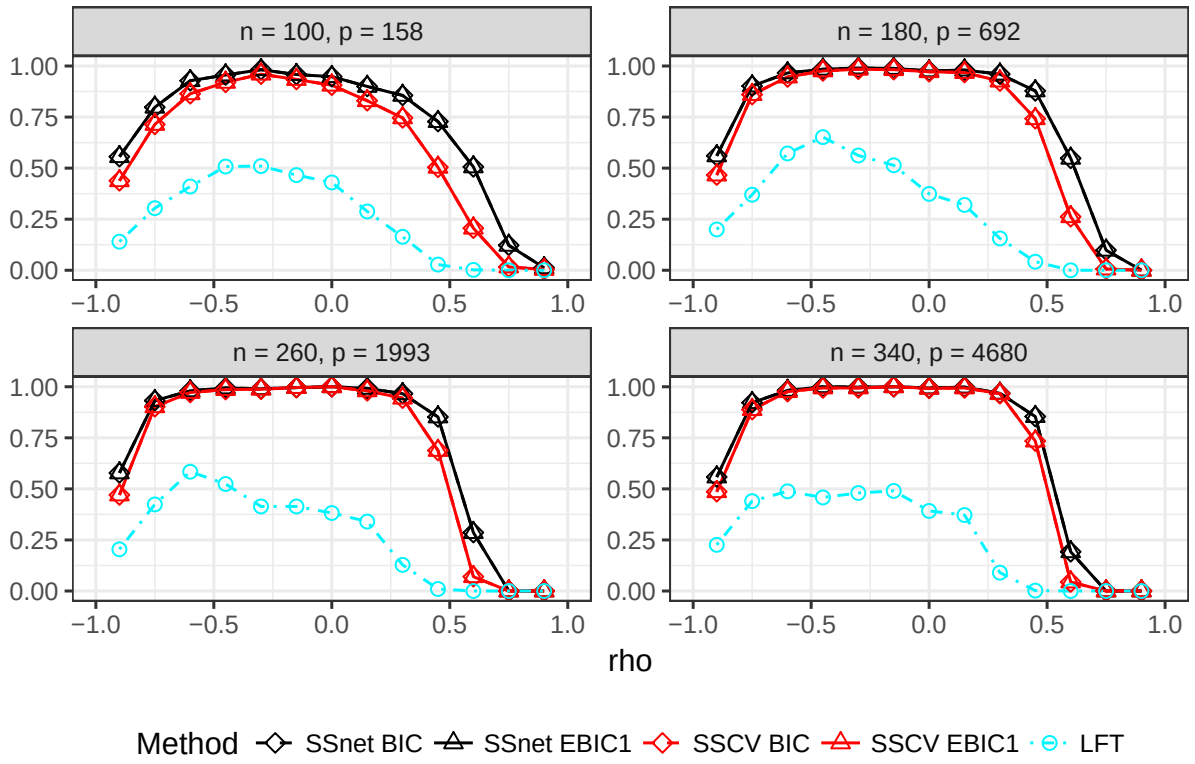


Figure 6.13: P_{inc} for models MF1-MF4 with $q = \Phi$

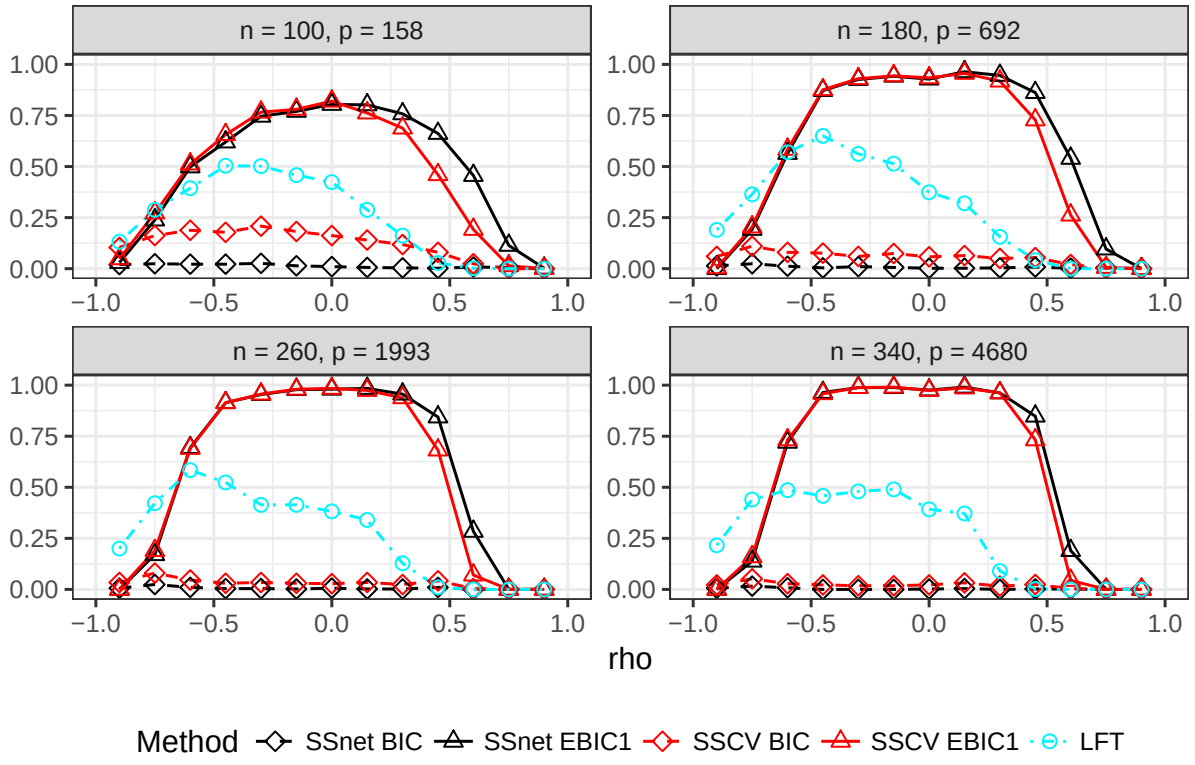


Figure 6.14: P_{equal} for models MF1-MF4 with $q = \Phi$

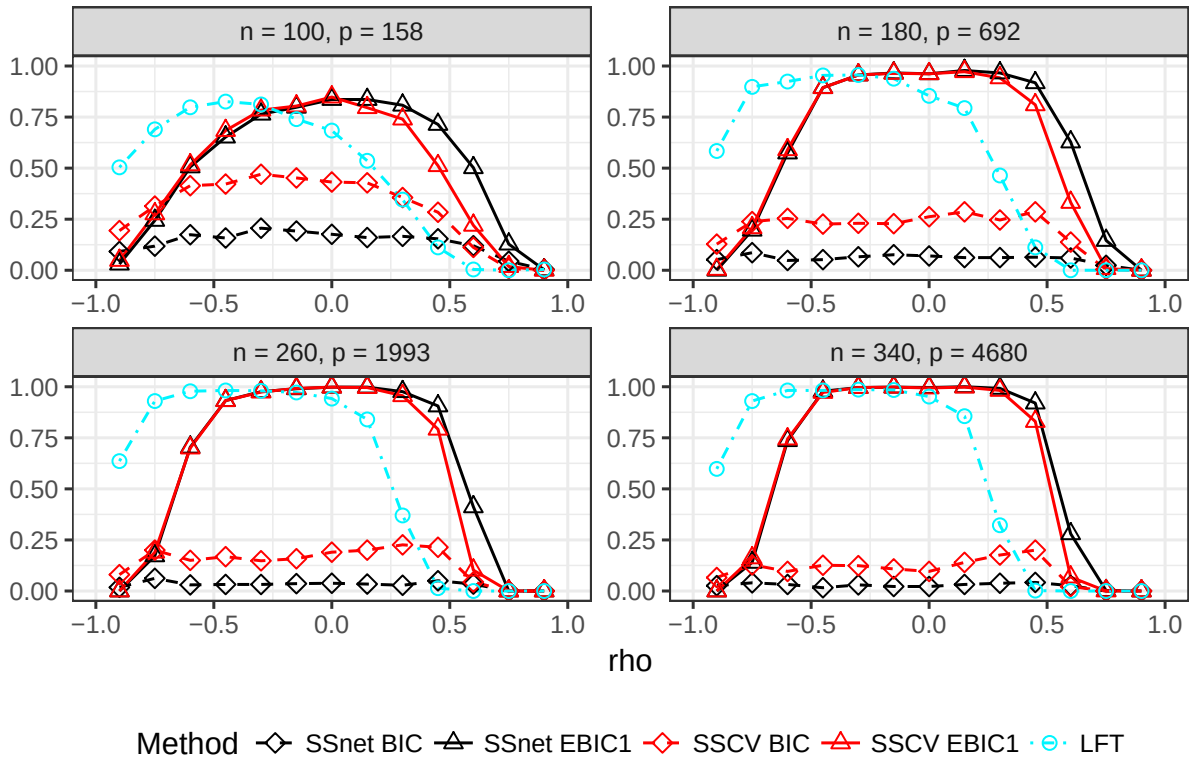


Figure 6.15: P_{supset} for models MF1-MF4 with $q = \Phi$

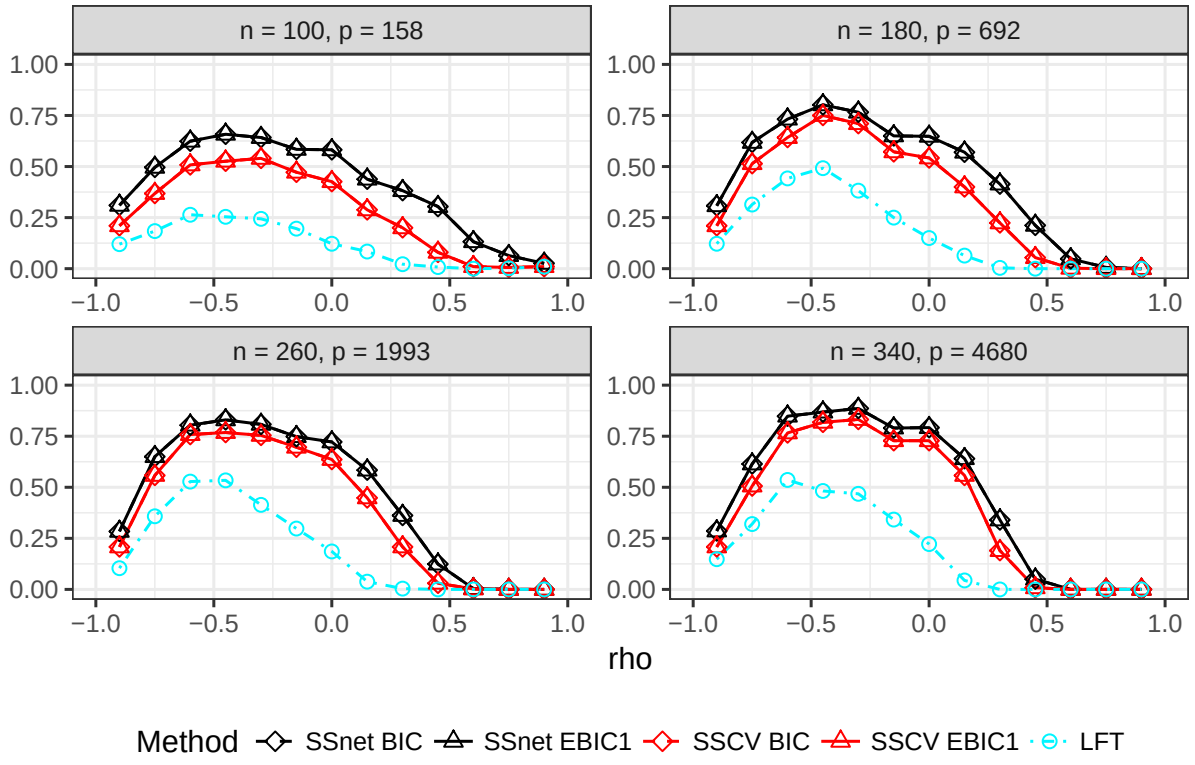


Figure 6.16: P_{inc} for models MF1-MF4 with $q = F_{Cauchy}$

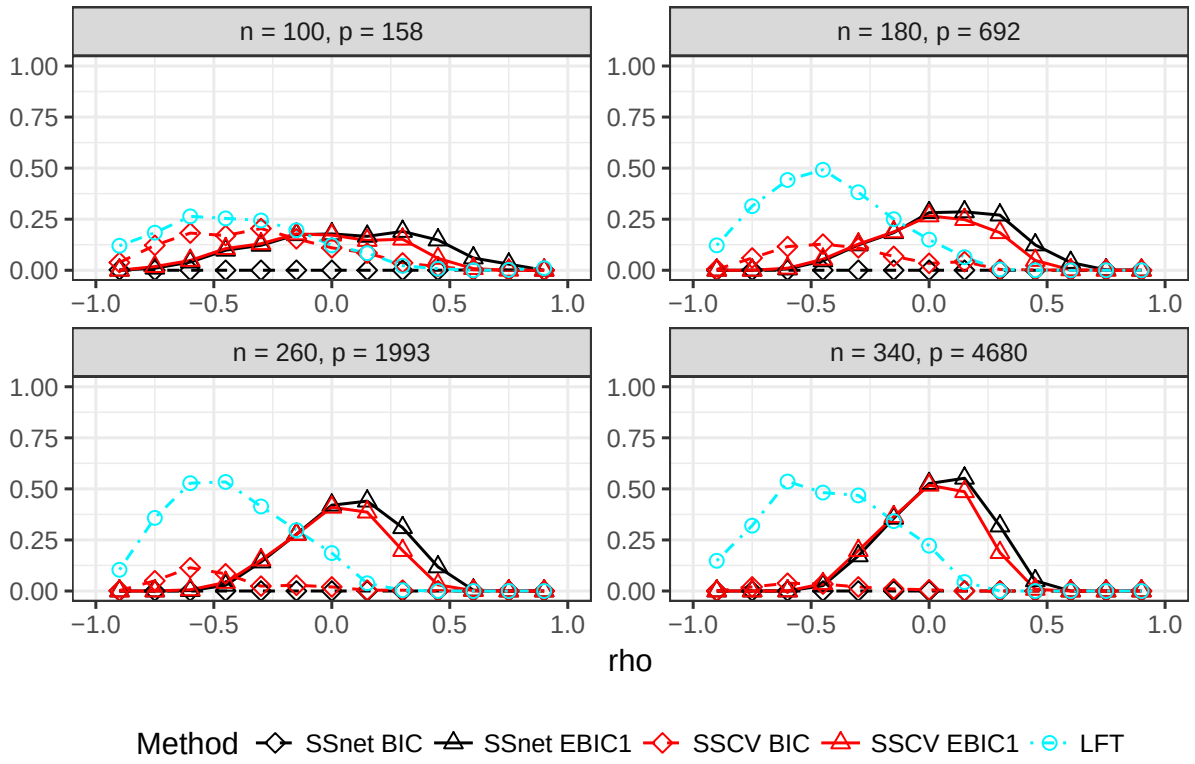


Figure 6.17: P_{equal} for models MF1-MF4 with $q = F_{Cauchy}$

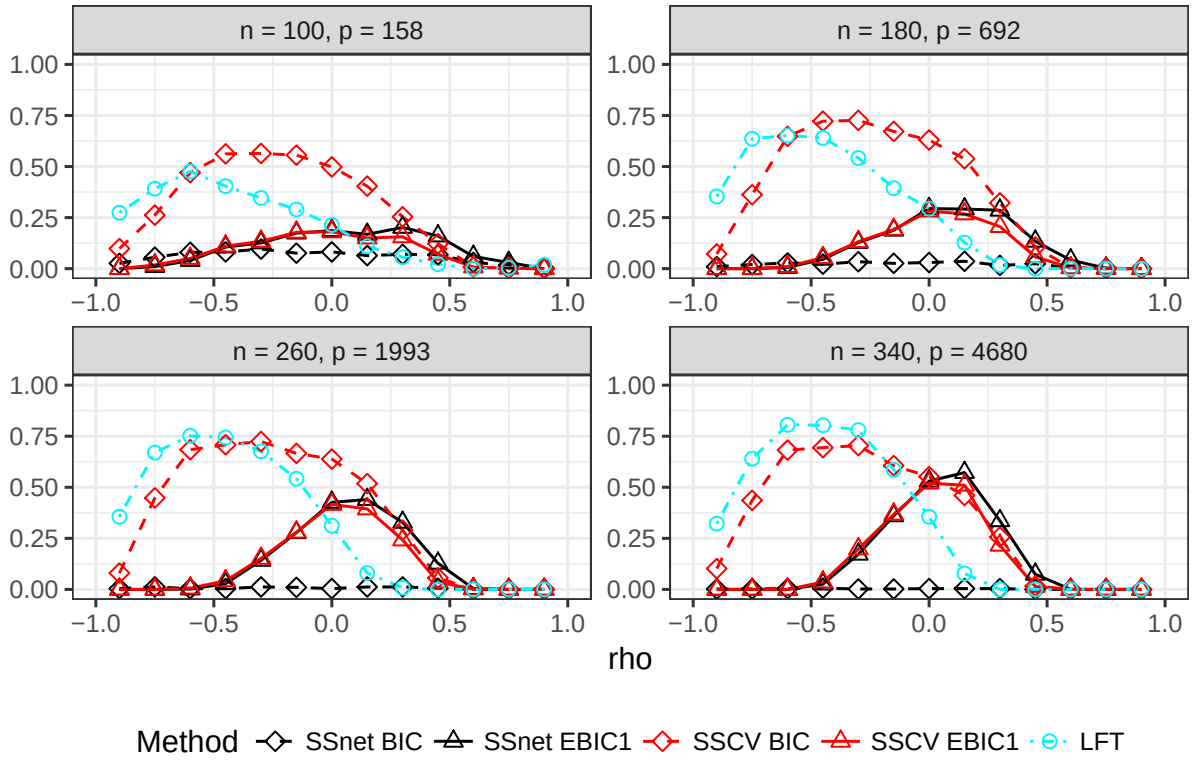


Figure 6.18: P_{supset} for models MF1-MF4 with $q = F_{Cauchy}$

Appendix A

Auxiliary definitions and lemmas

A.1. Existence and uniqueness of $\boldsymbol{\beta}^*$ for binary response

In this section we present results about existence and uniqueness of $\boldsymbol{\beta}^*$ defined in (1.3), when $Y \in \{0, 1\}$. The organization of this section is the following: Lemmas A.1-A.4 present auxiliary facts used in later part of this section. Lemma A.1 is used in Corollaries A.10-A.11 to show that risk function is strictly convex. Lemma A.2 is used in the proof of Remark A.13 and provides equivalent condition for positive-definiteness of covariance matrix $\boldsymbol{\Sigma}$, which can be easier to check than the condition that vector $\mathbf{X}$ is linearly nondegenerate. Lemma A.3 is a known fact in optimization (see Theorem 2.32 in Beck (2014)) which is crucial in the proof of Lemma A.4 and Theorem A.6. Lemma A.4 is a simple technical fact, which allows us to prove Theorem A.5 with the use of Lebesgue's monotone convergence theorem and without using Lebesgue's dominated convergence theorem. From Theorem A.5 follows existence and uniqueness of $\boldsymbol{\beta}^*$ also in the case of quadratic loss (see Remark A.13).

Theorems A.5-A.6 show that there exists minimum of risk function in any direction. This conclusion together with strict convexity of risk function is used in Lemma A.8 to prove that $\boldsymbol{\beta}^*$ exists (see Corollaries A.10-A.11). Then finally in Remarks A.12-A.14 we find sufficient conditions for existence and uniqueness of $\boldsymbol{\beta}^*$.

For original formulation of Theorem A.5 and A.8 see Li and Duan (1989). Note that the proof of Theorem A.5 is different from the proof in Li and Duan (1989), as we show directly in the proof how to avoid use of Lebesgue's dominated convergence theorem. Moreover, our proof of Lemma A.8 is different from the proof of Lemma 2.1 in Li and Duan (1989), because we show explicit way to construct a sequence described in Lemma 2.1 in Li and Duan (1989) and our construction uses one ball instead of two balls in $\mathbb{R}^{p_n+1}$ to prove the Lemma.

Lemma A.1. *If function* $g\colon \mathbb{R} \to \mathbb{R}$ *is strictly convex,* $\boldsymbol{X} \in \mathbb{R}^{p+1}$ *is a random vector, for all* $\boldsymbol{b} \in \mathbb{R}^{p+1}$*:* $\mathbb{E}|g(\boldsymbol{b}^T\boldsymbol{X})| < \infty$ *and for all* $\boldsymbol{b} \in \mathbb{R}^{p+1} \setminus \{\boldsymbol{0}_{p+1}\}$*:* $\mathbb{P}(\boldsymbol{b}^T\boldsymbol{X} = 0) < 1$*, then function* $f\colon \mathbb{R}^{p+1} \to \mathbb{R}$*, where* $f(\boldsymbol{b}) = \mathbb{E}g(\boldsymbol{b}^T\boldsymbol{X})$ *is strictly convex.*

Proof. Let $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^{p+1}$, $\mathbf{b}_1 \neq \mathbf{b}_2$ be some vectors and let $\alpha \in [0,1], \mathbf{b} = \alpha\mathbf{b}_1 + (1-\alpha)\mathbf{b}_2$. Let $A = \{\mathbf{b}_1^T\mathbf{X} = \mathbf{b}_2^T\mathbf{X}\}$. As $\mathbf{b}_1 \neq \mathbf{b}_2$, we have

$$\mathbb{P}(A) = \mathbb{P}(\mathbf{b}_1^T\mathbf{X} = \mathbf{b}_2^T\mathbf{X}) = \mathbb{P}((\mathbf{b}_1 - \mathbf{b}_2)^T\mathbf{X} = 0) < 1.$$

On set A^c we get from strict convexity of g:

$$g(\mathbf{b}^T\mathbf{X}) < \alpha g(\mathbf{b}_1^T\mathbf{X}) + (1-\alpha)g(\mathbf{b}_2^T\mathbf{X}).$$

We have, using convexity:

$$\begin{aligned} f(\mathbf{b}) &= \mathbb{E}g(\mathbf{b}^T\mathbf{X}) = \mathbb{E}g(\mathbf{b}^T\mathbf{X})I(A) + \mathbb{E}g(\mathbf{b}^T\mathbf{X})I(A^c) \\ &= \mathbb{E}(\alpha g(\mathbf{b}_1^T\mathbf{X}) + (1-\alpha)g(\mathbf{b}_2^T\mathbf{X}))I(A) + \mathbb{E}g(\mathbf{b}^T\mathbf{X})I(A^c) \\ &< \mathbb{E}(\alpha g(\mathbf{b}_1^T\mathbf{X}) + (1-\alpha)g(\mathbf{b}_2^T\mathbf{X}))I(A) \\ &\quad + \mathbb{E}(\alpha g(\mathbf{b}_1^T\mathbf{X}) + (1-\alpha)g(\mathbf{b}_2^T\mathbf{X}))I(A^c) = \alpha f(\mathbf{b}_1) + (1-\alpha)f(\mathbf{b}_2) \end{aligned}$$

as strict inequality follows from $\mathbb{P}(A^c) > 0$. □

Lemma A.2. *Let* $\boldsymbol{X} = (1, \tilde{\boldsymbol{X}}^T)^T \in \mathbb{R}^{p+1}$ *be a random vector satisfying* $\mathbb{E}||\tilde{\boldsymbol{X}}||_2^2 < \infty$*. Let* $\operatorname{Var}\tilde{\boldsymbol{X}} = \boldsymbol{\Sigma}$*. Then* $\boldsymbol{\Sigma} > 0$ *if and only if for every* $\boldsymbol{b} \in \mathbb{R}^{p+1} \setminus \{\boldsymbol{0}_{p+1}\}$*:* $\mathbb{P}(\boldsymbol{b}^T\boldsymbol{X} = 0) < 1$*.*

Proof. Condition $\boldsymbol{\Sigma} > 0$ is equivalent to

$$\forall \mathbf{b} = (b_0, \tilde{\mathbf{b}}^T)^T \in \mathbb{R}^{p+1}, \tilde{\mathbf{b}}^T \neq \mathbf{0}_p\colon\ 0 < \tilde{\mathbf{b}}^T\boldsymbol{\Sigma}\tilde{\mathbf{b}} = \operatorname{Var}(\tilde{\mathbf{b}}^T\tilde{\mathbf{X}}) = \operatorname{Var}(\mathbf{b}^T\mathbf{X}).$$

From this we obtain $\mathbb{P}(\mathbf{b}^T\mathbf{X} = 0) < 1$.

Now we need to prove that $\boldsymbol{\Sigma} > 0$ is implied by $\mathbb{P}(\mathbf{b}^T\mathbf{X} = 0) < 1$ for all $\mathbf{b} \in \mathbb{R}^{p+1} \setminus \{\mathbf{0}_{p+1}\}$. Firstly, we observe that

$$\forall \tilde{\mathbf{b}} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}\colon\ \mathbb{P}(\tilde{\mathbf{b}}^T\tilde{\mathbf{X}} = \mathbb{E}\tilde{\mathbf{b}}^T\tilde{\mathbf{X}}) = \mathbb{P}((-\mathbb{E}\tilde{\mathbf{b}}^T\tilde{\mathbf{X}}, \tilde{\mathbf{b}}^T)\mathbf{X} = 0) < 1.$$

Hence, we obtain:

$$\tilde{\mathbf{b}}^T\boldsymbol{\Sigma}\tilde{\mathbf{b}} = \operatorname{Var}(\tilde{\mathbf{b}}^T\tilde{\mathbf{X}}) = \mathbb{E}(\tilde{\mathbf{b}}^T\tilde{\mathbf{X}} - \mathbb{E}\tilde{\mathbf{b}}^T\tilde{\mathbf{X}})^2 = \mathbb{E}(\tilde{\mathbf{b}}^T\tilde{\mathbf{X}} - \mathbb{E}\tilde{\mathbf{b}}^T\tilde{\mathbf{X}})^2 I(\tilde{\mathbf{b}}^T\tilde{\mathbf{X}} \neq \mathbb{E}\tilde{\mathbf{b}}^T\tilde{\mathbf{X}}) > 0.$$

This means that $\boldsymbol{\Sigma} > 0$. □

Lemma A.3 (Beck (2014, Theorem 2.32))**.** *Let* $f\colon \mathbb{R}^n \to \mathbb{R}$ *be a continuous and coercive function, i.e.* $\lim_{||\boldsymbol{x}||_2 \to +\infty} f(\boldsymbol{x}) = +\infty$*. Let* $S \subseteq \mathbb{R}^n$ *be a nonempty closed set. Then* f *has a global minimum in* S*.*

Lemma A.4. *Let* $d\colon \mathbb{R} \to \mathbb{R}$ *be function satisfying* $\liminf_{t \to +\infty} d(t) > 0$ *and* $\limsup_{t \to -\infty} d(t) < 0$*. Let* $\tilde{R}$ *be continuous function such that for all* $s, t \in \mathbb{R}$ *we have:*

$$\tilde{R}(s) - \tilde{R}(t) \geq d(t)(s-t).$$

Then there exists

$$t^* = \arg\min_{t \in \mathbb{R}} \tilde{R}(t).$$

Proof. Because $\liminf_{t\to+\infty} d(t) > 0$, then there exist $t_1 \in \mathbb{R}, \eta_1 > 0$ such that for all $t \geq t_1$ $d(t) > \eta_1$. This means that for all $t \geq t_1$:

$$\tilde{R}(t) \geq \tilde{R}(t_1) + d(t_1)(t - t_1) > \tilde{R}(t_1) + \eta_1(t - t_1).$$

Hence $\lim_{t\to+\infty} \tilde{R}(t) = +\infty$. Analogously, from the fact that $\limsup_{t\to-\infty} d(t) < 0$ it follows that $\lim_{t\to-\infty} \tilde{R}(t) = +\infty$. This means that $\tilde{R}$ is continuous and coercive. By Lemma A.3, we obtain the conclusion of theorem. □

Theorem A.5 (based on Li and Duan (1989, Lemma 3.1 and Remark 3.2))**.** *Let* $\phi\colon\ \mathbb{R} \to \mathbb{R}$ *and* $g\colon\ \{0,1\} \to \mathbb{R}$ *be some functions. Define* $\rho\colon\ \mathbb{R} \times \{0,1\} \to \mathbb{R}$ *as*

$$\rho(b, y) = -yb + \phi(b) + g(y)$$

for all $b \in \mathbb{R}$ *and* $y \in \{0,1\}$ *be convex and differentiable function of* b*. Let* $\mathbb{E}||\boldsymbol{X}||_2 < \infty$*, for all* $\boldsymbol{b} \in \mathbb{R}^{p+1}$ $\mathbb{E}|\phi(\boldsymbol{b}^T\boldsymbol{X})| < \infty$ *and* $\mathbb{E}||\phi'(\boldsymbol{b}^T\boldsymbol{X})\boldsymbol{X}||_2 < \infty$*. Assume additionally that* $\mathbb{E}(Y|\boldsymbol{X}) \in (\phi'(-\infty), \phi'(+\infty))$ $\mathbb{P}_{\boldsymbol{X}}$ *a.e., where*

$$\phi'(+\infty) = \lim_{t\to+\infty} \phi'(t),\ \ \phi'(-\infty) = \lim_{t\to-\infty} \phi'(t).$$

Let for a given $\boldsymbol{b} \in \mathbb{R}^{p+1}$*:*

$$\tilde{R}(t) = R(t\boldsymbol{b}) = \mathbb{E}\rho(t\boldsymbol{b}^T\boldsymbol{X}, Y).$$

Then there exists $t^* \in \mathbb{R}$ *such that:*

$$t^* = \arg\min_{t\in\mathbb{R}} \tilde{R}(t).$$

Proof. Let $\tilde{l}(t, x, y) = l(t\mathbf{b}, x, y)$, where $l(\mathbf{b}, \mathbf{x}, y) = \rho(\mathbf{b}^T\mathbf{x}, y)$ and $d(t) = \mathbb{E}\frac{\partial\tilde{l}}{\partial t}(t, X, Y) = -\mathbb{E}Y\mathbf{b}^T\mathbf{X} + \mathbb{E}\phi'(t\mathbf{b}^T\mathbf{X})\mathbf{b}^T\mathbf{X}$. Function $\tilde{l}$ is convex and differentiable function of t, what follows from convexity and differentiability of ρ. Hence we obtain for all $s, t \in \mathbb{R}$ (see Theorem 25.1 in Rockafellar (1970)):

$$\tilde{l}(s, X, Y) - \tilde{l}(t, X, Y) \geq \frac{\partial\tilde{l}}{\partial t}(t, X, Y)(s - t).$$

Thus, after taking expectations, we get:

$$\tilde{R}(s) - \tilde{R}(t) \geq d(t)(s - t). \tag{A.1}$$

We observe that ϕ is convex and differentiable function, as we have from the definition of ρ: $\phi(b) = \rho(b, y) + yb - g(y)$. This means that ϕ' is nondecreasing function. Thus we obtain for all $t \geq s$ ($a_+ = aI(a > 0)$, $a_- = aI(a < 0)$):

$$\begin{aligned}\mathbb{E}Y\mathbf{b}^T\mathbf{X} + d(t) &= \mathbb{E}\phi'(t\mathbf{b}^T\mathbf{X})\mathbf{b}^T\mathbf{X} = \mathbb{E}\phi'(t\mathbf{b}^T\mathbf{X})(\mathbf{b}^T\mathbf{X})_+ + \mathbb{E}\phi'(t\mathbf{b}^T\mathbf{X})(\mathbf{b}^T\mathbf{X})_- \\ &\geq \mathbb{E}\phi'(s\mathbf{b}^T\mathbf{X})(\mathbf{b}^T\mathbf{X})_+ + \mathbb{E}\phi'(s\mathbf{b}^T\mathbf{X})(\mathbf{b}^T\mathbf{X})_- = \mathbb{E}\phi'(s\mathbf{b}^T\mathbf{X})\mathbf{b}^T\mathbf{X} = \mathbb{E}Y\mathbf{b}^T\mathbf{X} + d(s).\end{aligned}$$

In the above inequality we used the fact that if $\mathbf{b}^T\mathbf{X} < 0$, then $\phi'(t\mathbf{b}^T\mathbf{X}) \leq \phi'(s\mathbf{b}^T\mathbf{X})$ and thus

$$\phi'(t\mathbf{b}^T\mathbf{X})(\mathbf{b}^T\mathbf{X})_- \geq \phi'(s\mathbf{b}^T\mathbf{X})(\mathbf{b}^T\mathbf{X})_-.$$

Hence function d is nondecreasing. From the Lebesgue's monotone convergence theorem we have:

$$\lim_{t\to+\infty} d(t) + \mathbb{E}Y\boldsymbol{\beta}^T\mathbf{X} = \lim_{t\to+\infty} \mathbb{E}\phi'(t\boldsymbol{\beta}^T\mathbf{X})\boldsymbol{\beta}^T\mathbf{X} = \mathbb{E}\lim_{t\to+\infty} \phi'(t\boldsymbol{\beta}^T\mathbf{X})\boldsymbol{\beta}^T\mathbf{X}$$
$$= \mathbb{E}\phi'(+\infty)(\boldsymbol{\beta}^T\mathbf{X})_+ + \mathbb{E}\phi'(-\infty)(\boldsymbol{\beta}^T\mathbf{X})_- = \phi'(+\infty)\mathbb{E}(\boldsymbol{\beta}^T\mathbf{X})_+ + \phi'(-\infty)\mathbb{E}(\boldsymbol{\beta}^T\mathbf{X})_-,$$

$$\lim_{t\to-\infty} d(t) + \mathbb{E}Y\boldsymbol{\beta}^T\mathbf{X} = \lim_{t\to-\infty} \mathbb{E}\phi'(t\boldsymbol{\beta}^T\mathbf{X})\boldsymbol{\beta}^T\mathbf{X} = \mathbb{E}\lim_{t\to-\infty} \phi'(t\boldsymbol{\beta}^T\mathbf{X})\boldsymbol{\beta}^T\mathbf{X}$$
$$= \mathbb{E}\phi'(-\infty)(\boldsymbol{\beta}^T\mathbf{X})_+ + \mathbb{E}\phi'(+\infty)(\boldsymbol{\beta}^T\mathbf{X})_- = \phi'(-\infty)\mathbb{E}(\boldsymbol{\beta}^T\mathbf{X})_+ + \phi'(+\infty)\mathbb{E}(\boldsymbol{\beta}^T\mathbf{X})_-.$$

Thus, we get (as $\mathbb{E}(Y|\mathbf{X}) \in (\phi'(-\infty), \phi'(+\infty))$ $\mathbb{P}_{\mathbf{X}}$ a.e.):

$$\mathbb{E}Y\boldsymbol{\beta}^T\mathbf{X} = \mathbb{E}(\boldsymbol{\beta}^T\mathbf{X}\mathbb{E}(Y|\mathbf{X})) = \mathbb{E}((\boldsymbol{\beta}^T\mathbf{X})_+\mathbb{E}(Y|\mathbf{X})) + \mathbb{E}((\boldsymbol{\beta}^T\mathbf{X})_-\mathbb{E}(Y|\mathbf{X}))$$
$$< \phi'(+\infty)\mathbb{E}(\boldsymbol{\beta}^T\mathbf{X})_+ + \phi'(-\infty)\mathbb{E}(\boldsymbol{\beta}^T\mathbf{X})_- = \lim_{t\to+\infty} d(t) + \mathbb{E}Y\boldsymbol{\beta}^T\mathbf{X}.$$

This means that $\lim_{t\to+\infty} d(t) > 0$. Analogously, we get $\lim_{t\to-\infty} d(t) < 0$. Now, from the convexity of $\tilde{l}$ we have that function $\tilde{R}$ is convex. Because $\tilde{R}$ is convex function in open domain, it is continuous (see Roberts and Varberg (1973), chapter IV.41). Hence $\arg\min_{t\in\mathbb{R}} \tilde{R}(t)$ exists in view of Lemma A.4. □

Theorem A.6. *Let* $\pi\colon\ \mathbb{R} \to (0,1)$ *be nondecreasing function such that* $\ln\pi(b)$ *and* $\ln(1-\pi(b))$ *are concave functions of* b,

$$\lim_{b\to-\infty} \pi(b) = 0, \ \lim_{b\to+\infty} \pi(b) = 1.$$

Assume that $\boldsymbol{X} \in \mathbb{R}^{p+1}$ *is a random variable such that for all* $\boldsymbol{b} \in \mathbb{R}^{p+1} \setminus \{\boldsymbol{0}_{p+1}\}$ $\mathbb{P}(\boldsymbol{b}^T\boldsymbol{X} = 0) < 1$, $\mathbb{E}|\ln\pi(\boldsymbol{b}^T\boldsymbol{X})| < \infty$, $\mathbb{E}|\ln\Big(1-\pi(\boldsymbol{b}^T\boldsymbol{X})\Big)| < \infty$ *and* $Y \in \{0,1\}$ *is a random variable such that* $\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x}) = q(\boldsymbol{x}) \in (0,1)$ $\mathbb{P}_{\boldsymbol{X}}$ *a.e. Define:*

$$\rho(b,y) = -y\ln\pi(b) - (1-y)\ln(1-\pi(b)),$$

and let for a given $\boldsymbol{b} \in \mathbb{R}^{p+1}$

$$\tilde{R}(t) = R(t\boldsymbol{b}) = \mathbb{E}\rho(t\boldsymbol{b}^T\boldsymbol{X}, Y)$$

for $t \in \mathbb{R}$. *Then there exists* $t^* \in \mathbb{R}$ *such that:*

$$t^* = \arg\min_{t\in\mathbb{R}} \tilde{R}(t).$$

Proof. For a given $\mathbf{b} \neq \mathbf{0}_{p+1}$ we define set $A = \{\mathbf{b}^T\mathbf{X} = 0\}$. From our assumption $\mathbb{P}(A^c) > 0$. Observe that:

$$\lim_{t\to+\infty} \ln\pi(t\mathbf{b}^T\mathbf{X}) = \begin{cases} 0 & \mathbf{b}^T\mathbf{X} > 0 \\ -\infty & \mathbf{b}^T\mathbf{X} < 0 \\ \ln\pi(0) & \mathbf{b}^T\mathbf{X} = 0 \end{cases},$$

$$\lim_{t\to+\infty} \ln\Big(1-\pi(t\mathbf{b}^T\mathbf{X})\Big) = \begin{cases} -\infty & \mathbf{b}^T\mathbf{X} > 0 \\ 0 & \mathbf{b}^T\mathbf{X} < 0 \\ \ln(1-\pi(0)) & \mathbf{b}^T\mathbf{X} = 0 \end{cases}.$$

Thus on set $A^c \cap \{q(\mathbf{X}) \in (0,1)\}$ we have:

$$\lim_{t\to+\infty} q(\mathbf{X}) \ln \pi(t\mathbf{b}^T\mathbf{X}) + (1-q(\mathbf{X})) \ln\Big(1-\pi(t\mathbf{b}^T\mathbf{X})\Big) = -\infty.$$

Moreover, by conditioning on $\mathbf{X}$ we obtain:

$$R(\mathbf{b}) = -\mathbb{E}q(\mathbf{X}) \ln \pi(\mathbf{b}^T\mathbf{X}) - \mathbb{E}(1-q(\mathbf{X})) \ln\Big(1-\pi(\mathbf{b}^T\mathbf{X})\Big).$$

Hence from Lebesgue's monotone convergence theorem we obtain (as π is nondecreasing):

$$\lim_{t\to+\infty} \tilde{R}(t) = \mathbb{E}\rho(0,Y)I(A)$$
$$- \lim_{t\to+\infty} \mathbb{E}\Big(q(\mathbf{X}) \ln \pi(t\mathbf{b}^T\mathbf{X}) + (1-q(\mathbf{X})) \ln\Big(1-\pi(t\mathbf{b}^T\mathbf{X})\Big)\Big) I(A^c) = +\infty.$$

Analogously we obtain $\lim_{t\to-\infty} \tilde{R}(t) = +\infty$. Thus $\tilde{R}$ is coercive function. Moreover, $\tilde{R}$ is continuous, as it is convex function. Convexity of $\tilde{R}$ is implied by convexity of $\rho(\cdot, y)$ for all y. This means that existence of $t^* = \arg\min_{t\in\mathbb{R}} \tilde{R}(t)$ follows directly from Lemma A.3. □

Remark A.7. *Conditions* $\mathbb{E}|\ln \pi(\boldsymbol{b}^T\boldsymbol{X})| < \infty$, $\mathbb{E}|\ln\Big(1-\pi(\boldsymbol{b}^T\boldsymbol{X})\Big)| < \infty$ *in Theorem A.6 for logistic regression are satisfied for all* $\boldsymbol{b} \in \mathbb{R}^{p+1}$ *when* $\mathbb{E}||\boldsymbol{X}||_2 < \infty$. *See also Remark A.12.*

Lemma A.8 (based on Li and Duan (1989, Lemma 2.1))**.** *Let* $R\colon\ \mathbb{R}^{p+1} \to \mathbb{R}$ *for* $p \in \mathbb{N}$ *be strictly convex function satisfying the following property:*

$$\forall \boldsymbol{b} \in \mathbb{R}^{p+1}\ \exists t^* = \arg\min_{t\in\mathbb{R}} R(t\boldsymbol{b}).$$

Then there exists

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{p+1}} R(\boldsymbol{b}).$$

Proof. Because R is convex function in open domain, it is continuous (see Roberts and Varberg (1973), Chapter IV.41). Suppose that $\boldsymbol{\beta}^*$ does not exist. From continuity of R we can take a sequence $(\mathbf{b}_n)$ such that

$$R(\mathbf{b}_n) \to \inf_{\mathbf{b}\in\mathbb{R}^{p+1}} R(\mathbf{b}) \text{ and } ||\mathbf{b}_n||_2 \to \infty.$$

Let $\tilde{\mathbf{b}}_n = \mathbf{b}_n/||\mathbf{b}_n||_2$. Then $||\tilde{\mathbf{b}}_n||_2 = 1$. Moreover, set $S = \{\mathbf{b} \in \mathbb{R}^{p+1}\colon\ ||\mathbf{b}||_2 = 1\}$ is compact. This means that there exists subsequence $(\mathbf{b}_{k_n})$ such that $\tilde{\mathbf{b}}_{k_n} \to \tilde{\mathbf{b}}_0$ for some $\tilde{\mathbf{b}}_0 \in S$. From our assumption there exists

$$t_0 = \arg\min_{t\in\mathbb{R}} R(t\tilde{\mathbf{b}}_0).$$

Since R is strictly convex, this minimum is unique. Now we take $t_1 > t_0$. Then $R(t_1\tilde{\mathbf{b}}_0) > R(t_0\tilde{\mathbf{b}}_0)$ and from continuity of R there exists $\varepsilon > 0$ and $\eta > 0$ that for all $\mathbf{b} \in B(t_1\tilde{\mathbf{b}}_0, \varepsilon)$: $R(\mathbf{b}) > R(t_0\tilde{\mathbf{b}}_0) + \eta$.

Now let

$$\alpha_n = \frac{t_1 - t_0}{||\mathbf{b}_{k_n}||_2 - t_0}, \quad \mathbf{v}_n = \alpha_n\mathbf{b}_{k_n} + (1-\alpha_n)t_0\tilde{\mathbf{b}}_0.$$

We observe that $\alpha_n \to 0$ as $||\mathbf{b}_{k_n}||_2 \to \infty$ and

$$\lim_{n\to\infty} \mathbf{v}_n = \lim_{n\to\infty} \alpha_n\mathbf{b}_{k_n} + \lim_{n\to\infty}(1-\alpha_n)t_0\tilde{\mathbf{b}}_0 = \lim_{n\to\infty} \frac{t_1 - t_0}{||\mathbf{b}_{k_n}||_2 - t_0}||\mathbf{b}_{k_n}||_2\tilde{\mathbf{b}}_{k_n} + t_0\tilde{\mathbf{b}}_0$$

$$= (t_1 - t_0) \lim_{n\to\infty} \tilde{\mathbf{b}}_{k_n} + t_0\tilde{\mathbf{b}}_0 = (t_1 - t_0)\tilde{\mathbf{b}}_0 + t_0\tilde{\mathbf{b}}_0 = t_1\tilde{\mathbf{b}}_0.$$

This means that for large n we have $\mathbf{v}_n \in B(t_1\tilde{\mathbf{b}}_0, \varepsilon)$ and $R(\mathbf{v}_n) > R(t_0\tilde{\mathbf{b}}_0) + \eta$. From strict convexity for large n we have:

$$R(t_0\tilde{\mathbf{b}}_0) + \eta < R(\mathbf{v}_n) < \alpha_n R(\mathbf{b}_{k_n}) + (1 - \alpha_n)R(t_0\tilde{\mathbf{b}}_0).$$

Hence after simple transformations we obtain $R(\mathbf{b}_{k_n}) > R(t_0\tilde{\mathbf{b}}_0) + \frac{\eta}{\alpha_n}$, but from the property

$$R(\mathbf{b}_{k_n}) \to \inf_{\mathbf{b}\in\mathbb{R}^{p+1}} R(\mathbf{b})$$

we get a contradiction as $\alpha_n \to 0$. Hence $\boldsymbol{\beta}^*$ exists. □

Example A.9. *Assumption of strict convexity of R in Lemma A.8 cannot be omitted. Let $\boldsymbol{b} = (x, y)^T$ and consider the function $R(\boldsymbol{b}) = R(x, y) = \max(x, y + x^2)$. We define $g_{(x,y)}(t) = R(tx, ty)$. Function R is convex as a maximum of convex functions. We will show that for all $(x, y) \in \mathbb{R}^2$ function $g_{(x,y)}$ has a minimum. We consider 3 cases:*
Case 1: $(x, y) = (0, 0)$.
In this case $g_{(x,y)}(t) = 0$ for all $t \in \mathbb{R}$ and thus it has a minimum.
Case 2: $x \neq 0$.
In this case $g_{(x,y)}(t) = t^2x^2 + ty$ for $|t| > (x - y)/x^2$, thus

$$\lim_{t\to\pm\infty} g_{(x,y)}(t) = +\infty.$$

This property and convexity of $g_{(x,y)}$ imply that $g_{(x,y)}$ has a minimum (see Lemma A.3).
Case 3: $x = 0, y \neq 0$.
In this case $g_{(x,y)}(t) = \max(0, ty) \geq 0 = g_{(x,y)}(0)$. Thus $g_{(x,y)}$ has a minimum.
Function R does not have a minimum, as we have $R(x, -x^2 + x) = \max(x, x) = x$.

Corollary A.10. *Assume that assumptions of Theorem A.5 hold, $\rho(\cdot, y)$ is strictly convex function for all y and for all $\boldsymbol{b} \in \mathbb{R}^{p+1} \setminus \{\boldsymbol{0}_{p+1}\}$: $\mathbb{P}(\boldsymbol{b}^T\boldsymbol{X} = 0) < 1$. Then there exists unique*

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{p+1}} R(\boldsymbol{b}).$$

Proof. The proof follows directly from Theorem A.5 and Lemma A.8 after noting that for all $\mathbf{b} \in \mathbb{R}^{p+1} \setminus \{\mathbf{0}_{p+1}\}$ the function $\tilde{R}(t) = \mathbb{E}l(t\mathbf{b}, \mathbf{X}, Y)$ is strictly convex in view of strict convexity of ρ and Lemma A.1. □

Corollary A.11. *Assume that assumptions of Theorem A.6 hold and $\ln \pi(b)$, $\ln(1 - \pi(b))$ are strictly concave functions of b. Then there exists unique*

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{p+1}} R(\boldsymbol{b}).$$

Proof. The proof follows directly from Theorem A.6 and Lemma A.8 after noting that for all $\mathbf{b} \in \mathbb{R}^{p+1} \setminus \{\mathbf{0}_{p+1}\}$ the function $\tilde{R}(t) = \mathbb{E}l(t\mathbf{b}, \mathbf{X}, Y)$ is strictly convex in view of strict convexity of ρ (what follows from strict concavity of $\ln \pi(b)$ and $\ln(1 - \pi(b))$) and Lemma A.1. □

We prove the following remark using Corollary A.10, but the same conclusions can be obtained from Corollary A.11.

Remark A.12. *Unique $\boldsymbol{\beta}^*$ exists for logistic loss:*

$$l(\boldsymbol{b}, \boldsymbol{x}, y) = -y\boldsymbol{x}^T\boldsymbol{b} + \ln\Big(1 + \exp\Big(\boldsymbol{x}^T\boldsymbol{b}\Big)\Big)$$

if the following conditions are satisfied: $\mathbb{E}||\boldsymbol{X}||_2 < \infty$, *for all* $\boldsymbol{b} \in \mathbb{R}^{p+1} \setminus \{\boldsymbol{0}_{p+1}\}$*:* $\mathbb{P}(\boldsymbol{b}^T\boldsymbol{X} = 0) < 1$ *and* $\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) \in (0, 1)$ $\mathbb{P}_{\boldsymbol{X}}$ *a.e.*

Proof. We have $g(y) \equiv 0$, $\phi(b) = \ln\Big(1 + e^b\Big)$, $\phi'(b) = (1 + e^{-b})^{-1}$. Function ϕ is strictly convex, therefore l is strictly convex with respect to $\mathbf{b}$. Obviously, l is continuously differentiable.
Now we observe that $|\phi'(b)| \leq 1$ thus ϕ is a Lipschitz function and we have:

$$\mathbb{E}|\phi(\mathbf{b}^T\mathbf{X})| \leq \mathbb{E}|\phi(\mathbf{b}^T\mathbf{X}) - \phi(0)| + |\phi(0)| \leq \mathbb{E}|\mathbf{b}^T\mathbf{X}| + \ln 2 \leq ||\mathbf{b}||_2\mathbb{E}||\mathbf{X}||_2 + \ln 2 < \infty,$$
$$\mathbb{E}||\phi'(\mathbf{b}^T\mathbf{X})\mathbf{X}||_2 = \mathbb{E}|\phi'(\mathbf{b}^T\mathbf{X})|||\mathbf{X}_2||_2 \leq \mathbb{E}||\mathbf{X}||_2 < \infty.$$

We see that $\phi'(-\infty) = 0$, $\phi'(+\infty) = 1$ and $\mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) \in (0, 1)$. Hence $\boldsymbol{\beta}^*$ exists in view of Corollary A.10. □

Remark A.13. *Unique $\boldsymbol{\beta}^*$ exists for quadratic loss:*

$$l(\boldsymbol{b}, \boldsymbol{x}, y) = \frac{1}{2}(y - \boldsymbol{x}^T\boldsymbol{b})^2 = -y\boldsymbol{x}^T\boldsymbol{b} + \frac{1}{2}(\boldsymbol{x}^T\boldsymbol{b})^2 + \frac{1}{2}y^2$$

if $\mathbb{E}||\boldsymbol{X}||_2^2 < \infty$ *and* $\operatorname{Var}\tilde{\boldsymbol{X}} = \boldsymbol{\Sigma} > 0$.

Proof. We have $g(y) = y^2/2$, $\phi(b) = b^2/2$, $\phi'(b) = b$. Function ϕ is a strictly convex function, therefore l is strictly convex with respect to $\mathbf{b}$. Obviously, l is differentiable with respect to $\mathbf{b}$.
Condition $\mathbb{E}||\mathbf{X}||_2 < \infty$ follows from $\mathbb{E}||\mathbf{X}||_2^2 < \infty$. Then we check moment conditions:

$$\mathbb{E}|\phi(\mathbf{b}^T\mathbf{X})| = \frac{1}{2}\mathbb{E}|\mathbf{b}^T\mathbf{X}|^2 \leq \frac{1}{2}||\mathbf{b}||_2^2\mathbb{E}||\mathbf{X}||_2^2 < \infty,$$
$$\mathbb{E}||\phi'(\mathbf{b}^T\mathbf{X})\mathbf{X}||_2 = \mathbb{E}|\mathbf{b}^T\mathbf{X}|||\mathbf{X}||_2 \leq ||\mathbf{b}||_2\mathbb{E}||\mathbf{X}||_2^2 < \infty.$$

$\mathbb{P}(\mathbf{b}^T\mathbf{X} = 0) < 1$ follows from $\boldsymbol{\Sigma} > 0$ in view of Lemma A.2.
Moreover, we have: $\mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) \in [0, 1] \subset \mathbb{R} = (\phi'(-\infty), \phi'(+\infty))$.
Hence existence of $\boldsymbol{\beta}^*$ follows from Corollary A.10. □

Remark A.14. *Unique $\boldsymbol{\beta}^*$ exists for probit loss:*

$$l(\boldsymbol{b}, \boldsymbol{x}, y) = -y\ln\Phi(\boldsymbol{x}^T\boldsymbol{b}) - (1 - y)\ln\Big(1 - \Phi(\boldsymbol{x}^T\boldsymbol{b})\Big),$$

if the following conditions are satisfied: $\mathbb{E}||\boldsymbol{X}||_2^2 < \infty$, $\operatorname{Var}\boldsymbol{X} = \boldsymbol{\Sigma} > 0$ *and* $\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) \in (0, 1)$ $\mathbb{P}_{\boldsymbol{X}}$ *a.e.*

Proof. Firstly, we observe that $\log\Phi(x)$ is strictly concave function, as we have:

$$(\log\Phi(x))'' = \frac{\phi(x)(-x\Phi(x) - \phi(x))}{\Phi^2(x)} < 0,$$

where $\phi(x) = \Phi'(x) > 0$ and $x\Phi(x) + \phi(x) > 0$ what follows from Mill's inequality (see Gordon (1941)):

$$\frac{x}{x^2+1} < \frac{1-\Phi(x)}{\phi(x)} < \frac{1}{x}.$$

Analogously $\log(1-\Phi(x)) = \log\Phi(-x)$ is strictly concave function. $\pi(s) = \Phi(s)$ is a cdf of $\mathcal{N}(0,1)$ hence the assumptions of Theorem A.6 regarding monotonicity and limits in $\pm\infty$ of Φ are satisfied. From Birnbaum's inequality (see Birnbaum (1942)) valid for $x > 0$:

$$1 - \Phi(x) > \frac{2\phi(x)}{\sqrt{x^2+4}+x}$$

and from inequality $\sqrt{x^2+4} \leq x+2$ for $x > 0$ we have:

$$\begin{aligned}
\mathbb{E}|\ln\Big(1-\Phi(\mathbf{b}^T\mathbf{X})\Big)| &= -\mathbb{E}\ln\Big(1-\Phi(\mathbf{b}^T\mathbf{X})\Big) = -\mathbb{E}\ln\Big(1-\Phi(\mathbf{b}^T\mathbf{X})\Big)I(\mathbf{b}^T\mathbf{X}>0) \\
&\quad - \mathbb{E}\ln\Big(1-\Phi(\mathbf{b}^T\mathbf{X})\Big)I(\mathbf{b}^T\mathbf{X}\leq 0) \\
&\leq -\mathbb{E}\ln\Big(2\phi(\mathbf{b}^T\mathbf{X})\Big)I(\mathbf{b}^T\mathbf{X}>0) \\
&\quad + \mathbb{E}\ln\Big(\sqrt{(\mathbf{b}^T\mathbf{X})^2+4}+\mathbf{b}^T\mathbf{X}\Big)I(\mathbf{b}^T\mathbf{X}>0) + \mathbb{P}(\mathbf{b}^T\mathbf{X}\leq 0)\ln 2 \\
&\leq -\mathbb{P}(\mathbf{b}^T\mathbf{X}>0)\ln 2 + \frac{1}{2}\mathbb{E}(\mathbf{b}^T\mathbf{X})^2 I(\mathbf{b}^T\mathbf{X}>0) + \mathbb{P}(\mathbf{b}^T\mathbf{X}>0)\ln\sqrt{2\pi} \\
&\quad + \mathbb{E}(\sqrt{(\mathbf{b}^T\mathbf{X})^2+4}+\mathbf{b}^T\mathbf{X}-1)I(\mathbf{b}^T\mathbf{X}>0) + \ln 2 \\
&\leq \frac{1}{2}\mathbb{E}(\mathbf{b}^T\mathbf{X})^2 I(\mathbf{b}^T\mathbf{X}>0) + \mathbb{P}(\mathbf{b}^T\mathbf{X}>0)\ln\sqrt{\frac{\pi}{2}} \\
&\quad + \mathbb{E}(\sqrt{(\mathbf{b}^T\mathbf{X})^2+4\mathbf{b}^T\mathbf{X}+4}+\mathbf{b}^T\mathbf{X}-1)I(\mathbf{b}^T\mathbf{X}>0) + \ln 2 \\
&\leq 2\ln 2 + \frac{1}{2}||\mathbf{b}||_2^2\mathbb{E}||\mathbf{X}||_2^2 + 2||\mathbf{b}||_2\mathbb{E}||\mathbf{X}||_2 + 1 < \infty.
\end{aligned}$$

Analogously we obtain $\mathbb{E}|\ln\Phi(\mathbf{b}^T\mathbf{X})| = \mathbb{E}|\ln\Big(1-\Phi(-\mathbf{b}^T\mathbf{X})\Big)| < \infty$. This means that $\boldsymbol{\beta}^*$ exists and is unique in view of Corollary A.11. □

A.2. Elliptically contoured distributions

Main aim of this section is to discuss basic properties of elliptically contoured distributions and their relation with linear regressions condition, which is used in Chapters 2-3.

Definition A.15. *We say that random vector $\boldsymbol{X} \in \mathbb{R}^p$, where $p \in \mathbb{N}$ follows elliptically contoured distribution with parameters $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, where $\boldsymbol{\Sigma}$ is nonnegative definite matrix ($\boldsymbol{\Sigma} \geq 0$) if characteristic function of $\boldsymbol{X}$ is of the form $\psi_{\boldsymbol{X}}(\boldsymbol{t}) = e^{i\boldsymbol{t}^T\boldsymbol{\mu}}\phi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^p$ and some $\phi\colon\ \mathbb{R} \to \mathbb{C}$. In this case we write $\boldsymbol{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$.*

Definition A.16. *We say that random vector $\boldsymbol{X} \in \mathbb{R}^p$, where $p \in \mathbb{N}$ follows spherically contoured distribution if characteristic function of $\boldsymbol{X}$ is of the form $\psi_{\boldsymbol{X}}(\boldsymbol{t}) = \phi(\boldsymbol{t}^T\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^p$ and some $\phi\colon\ \mathbb{R} \to \mathbb{C}$. In this case we write $\boldsymbol{X} \sim SC_p(\phi)$.*

Remark A.17. *If* $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *then* $\boldsymbol{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$, *where* $\phi(s) = e^{-\frac{s}{2}}$.

Theorem A.18 (Cambanis et al. (1981, Theorem 1))**.** $\boldsymbol{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ *with* $\operatorname{rank} \boldsymbol{\Sigma} = k$, *where* $k \leq p$ *if and only if*

$$\boldsymbol{X} \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{A}\,\boldsymbol{U}^{(k)} R,$$

where $R \geq 0$ *is independent of* $\boldsymbol{U}^{(k)}$, $\boldsymbol{U}^{(k)}$ *is uniformly distributed on the unit sphere in* $\mathbb{R}^k$, $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$ *is a rank factorization of* $\boldsymbol{\Sigma}$ *(*$\boldsymbol{A} \in \mathbb{R}^{p \times k}$, $\operatorname{rank} \boldsymbol{A} = k$*), and the distribution function* F *of* R *is related to* ϕ *as follows:*

$$\phi(u) = \int\limits_{[0,\infty)} \Omega_k(r^2 u)\; dF(r),$$

where $u \geq 0$, $\Omega_k(\boldsymbol{t}) := \Omega_k(||\boldsymbol{t}||^2)$ *(*$\boldsymbol{t} \in \mathbb{R}^k$*) is the characteristic function of* $\boldsymbol{U}^{(k)}$.

Theorem A.19 (Cambanis et al. (1981, Corollary 5))**.** *Let*

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{A}\,\boldsymbol{U}^{(k)} R \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$$

with $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}$ *and* $\operatorname{rank} \boldsymbol{A} = \operatorname{rank} \boldsymbol{\Sigma} = k \geq 1$. *Further, let*

$$\boldsymbol{X} = (\boldsymbol{X}_1^T, \boldsymbol{X}_2^T)^T, \quad \boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where $\boldsymbol{X}_1, \boldsymbol{\mu}_1 \in \mathbb{R}^m$, $\Sigma_{11} \in \mathbb{R}^{m \times m}$, *and assume* $k_2 = \operatorname{rank} \boldsymbol{\Sigma}_{22} \geq 1$ *and* $k_1 = k - k_2 \geq 1$. *Finally let* S *denote the column space of* $\boldsymbol{\Sigma}_{22}$. *Then a regular conditional distribution of* $\boldsymbol{X}$, *given* $\boldsymbol{X}_2 = \boldsymbol{x}_2$, *is given by:*

$$\begin{aligned} &(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) \sim EC_m(\boldsymbol{\mu}_{\boldsymbol{x}_2}, \boldsymbol{\Sigma}^*, \phi_{d(\boldsymbol{x}_2)}) && \text{for } \boldsymbol{x}_2 \in \boldsymbol{\mu}_2 + S, \\ &(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) \stackrel{d}{=} \boldsymbol{\mu}_1 && \text{for } \boldsymbol{x}_2 \notin \boldsymbol{\mu}_2 + S, \end{aligned}$$

with a full rank representation

$$(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) \stackrel{d}{=} \boldsymbol{\mu}_{\boldsymbol{x}_2} + \boldsymbol{A}^* \boldsymbol{U}^{(k_1)} R_{d(\boldsymbol{x}_2)} \text{ for } \boldsymbol{x}_2 \in \boldsymbol{\mu}_2 + S,$$

where $\boldsymbol{\mu}_{\boldsymbol{x}_2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$, $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-\boldsymbol{\Sigma}_{21}$, $d(\boldsymbol{x}_2) = (\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^- (\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$, $\boldsymbol{\Sigma}_{22}^-$ *is a generalized inverse of* $\boldsymbol{\Sigma}_{22}$ *and* $\boldsymbol{\Sigma}^* = \boldsymbol{A}^* \boldsymbol{A}^{*T}$ *is a rank factorization of* $\boldsymbol{\Sigma}^*$ *and* $\operatorname{rank} \boldsymbol{A}^* = k_1$. *Moreover,* $R_{d(\boldsymbol{x}_2)}$ *is independent of* $\boldsymbol{U}^{(k_1)}$.

Corollary A.20. *Let* $\boldsymbol{X} \in \mathbb{R}^p$ *be a random vector such that* $\mathbb{E}||\boldsymbol{X}||_2 < \infty$ *and assumptions of the Theorem A.19 are satisfied. Then for all* $\boldsymbol{\beta} \in \mathbb{R}^p$ *such that* $\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} \neq 0$ *we have for all* $z \in \mathbb{R}$*:*

$$\mathbb{E}(\boldsymbol{X} | \boldsymbol{X}^T \boldsymbol{\beta} = z) = \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta})^{-1}(z - \boldsymbol{\mu}^T \boldsymbol{\beta}).$$

Proof. Since $\mathbf{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$, then we have $(\mathbf{X}^T, \mathbf{X}^T\boldsymbol{\beta})^T \sim EC_{p+1}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \phi)$, where

$$\tilde{\boldsymbol{\mu}} = (\boldsymbol{\mu}^T, \boldsymbol{\mu}^T\boldsymbol{\beta})^T \text{ and } \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{\beta} \\ \boldsymbol{\beta}^T\boldsymbol{\Sigma} & \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \end{bmatrix}.$$

To prove this statement, we observe that:

$$\psi_{(\mathbf{X}^T, \mathbf{X}^T\boldsymbol{\beta})^T}((\mathbf{t}^T, s)^T) = \mathbb{E}e^{\mathbf{t}^T\mathbf{X} + s\mathbf{X}^T\boldsymbol{\beta}} = \mathbb{E}e^{\mathbf{X}^T(\mathbf{t} + s\boldsymbol{\beta})} = \psi_{\mathbf{X}}(\mathbf{t} + s\boldsymbol{\beta})$$

$$= e^{i(\mathbf{t}+s\boldsymbol{\beta})^T\boldsymbol{\mu}}\phi((\mathbf{t}+s\boldsymbol{\beta})^T\boldsymbol{\Sigma}(\mathbf{t}+s\boldsymbol{\beta})) = e^{i(\mathbf{t}^T,s)\tilde{\boldsymbol{\mu}}}\phi((\mathbf{t}^T,s)\tilde{\boldsymbol{\Sigma}}(\mathbf{t}^T,s)^T).$$

Hence from Theorem A.19 we obtain for $z \in \mathbb{R}$:

$$(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta} = z)^T \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}(z - \boldsymbol{\mu}^T\boldsymbol{\beta}) + \mathbf{A}^*\mathbf{U}^{(k)}R_{d(z)}.$$

Because $\mathbb{E}||\mathbf{X}||_2 < \infty$ and $\mathbb{E}\mathbf{U}^{(k)} = 0$ and $\mathbf{U}^{(k)}$ and $R_{d(z)}$ are independent, we get:

$$\mathbb{E}(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta} = z) = \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}(z - \boldsymbol{\mu}^T\boldsymbol{\beta}).$$

□

Proof of the following Corollary is identical to proof of Corollary A.20.

Corollary A.21. *Let $\boldsymbol{X} \in \mathbb{R}^p$ be a random vector such that $\mathbb{E}||\boldsymbol{X}||_2 < \infty$ and all of the assumptions of the Theorem A.19 are satisfied. If $\boldsymbol{\Sigma} > 0$ and $\boldsymbol{B} \in \mathbb{R}^{p\times k}$ is a matrix such that* $\operatorname{rank}\boldsymbol{B} = k$. *Then for all $\boldsymbol{z} \in \mathbb{R}^k$ we have:*

$$\mathbb{E}(\boldsymbol{X}|\boldsymbol{X}^T\boldsymbol{B} = \boldsymbol{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}(\boldsymbol{z} - \boldsymbol{B}^T\boldsymbol{\mu}).$$

Now we want to characterize distributions of $\mathbf{X}$ satisfying conclusion of Corollary A.21. Lemma A.22 is a basic tool here, as it allows to characterize spherically contoured distributions and is used in Theorem A.23 to characterize elliptically contoured distributions (see also Hardin (1982) for similar results).

Lemma A.22 (Eaton (1986, Theorem 1))**.** *Suppose the random vector $\boldsymbol{X} \in \mathbb{R}^p$ satisfies $\mathbb{E}||\boldsymbol{X}||_2 < \infty$. Assume that for each vector $\boldsymbol{v} \neq \boldsymbol{0}_p$ and for each vector $\boldsymbol{u}$ which is perpendicular to $\boldsymbol{v}$ (that is $\boldsymbol{u}^T\boldsymbol{v} = 0$),*

$$\mathbb{E}(\boldsymbol{u}^T\boldsymbol{X}|\boldsymbol{v}^T\boldsymbol{X}) = 0. \tag{A.2}$$

Then $\boldsymbol{X}$ is spherically contoured and conversely, if $\boldsymbol{X}$ is spherically contoured, then (A.2) is satisfied.

Theorem A.23. *Suppose the random vector $\boldsymbol{X} \in \mathbb{R}^p$ ($p \geq 2$) satisfies $\mathbb{E}||\boldsymbol{X}||_2 < \infty$, $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}$. Assume that exists $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Sigma} > 0$ and $k \in \{1, \ldots, p-1\}$ such that for all $\boldsymbol{B} \in \mathbb{R}^{p\times k}$ with* $\operatorname{rank}\boldsymbol{B} = k$ *the following equality holds:*

$$\mathbb{E}(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}) = (\boldsymbol{I}_p - \boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(\tilde{\boldsymbol{B}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{B}}^T)\boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{X}. \tag{A.3}$$

Then $\boldsymbol{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ for some function $\phi\colon\ \mathbb{R} \to \mathbb{C}$.

Proof. From the Lemma 3.19 we get for $\mathbf{Z} = \mathbf{X} - \boldsymbol{\mu}$:

$$\mathbb{E}(\mathbf{Z}|\mathbf{B}^T\mathbf{Z}) = \boldsymbol{\Sigma}\mathbf{B}(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Z}. \tag{A.4}$$

As $\boldsymbol{\Sigma}$ is invertible because it is positive definite, we define $\mathbf{V} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Z}$ and $\mathbf{C} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{B}$. After substitution of this equalities into (A.4) we obtain for every $\mathbf{C} \in \mathbb{R}^{p\times k}$ with $\operatorname{rank}\mathbf{C} = k$:

$$\mathbb{E}(\mathbf{V}|\mathbf{C}^T\mathbf{V}) = \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{V}. \tag{A.5}$$

Now we will prove that $\mathbf{V}$ follows spherically contoured distribution. Let $\mathbf{b}_1 = \mathbf{b} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$ be any vector and let $\mathbf{b}_{k+1} = \mathbf{w} \in \mathbb{R}^p$ be vector perpendicular to $\mathbf{b}$. We can find vectors $\mathbf{b}_2, \ldots, \mathbf{b}_k \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$ such that all vectors $\mathbf{b}_i$ for $i = 1, \ldots, k+1$ are orthogonal as $1 \leq k < p$ (for example by Gram-Schmidt orthogonalization). Let $\mathbf{C} = [\mathbf{b}_1, \ldots, \mathbf{b}_k]$. Then we have in view of (A.5):

$$\mathbb{E}(\mathbf{w}^T\mathbf{V}|\mathbf{b}^T\mathbf{V}) = \mathbb{E}(\mathbb{E}(\mathbf{w}^T\mathbf{V}|\mathbf{C}^T\mathbf{V})|\mathbf{b}^T\mathbf{V}) = \underbrace{\mathbf{w}^T\mathbf{C}}_{\mathbf{0}_k^T}(\mathbf{C}^T\mathbf{C})^{-1}\mathbb{E}(\mathbf{C}^T\mathbf{V}|\mathbf{b}^T\mathbf{V}) = 0.$$

This means that $\mathbf{V}$ follows spherically contoured distribution in view of Lemma A.22 and, consequently, $\mathbf{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{V} + \boldsymbol{\mu}$ follows elliptically contoured distribution. □

A.3. Existence, sparseness and uniqueness of $\hat{\beta}_L$

Facts presented in this section concern model without intercept, but they can be easily generalized to the case of the model with intercept. The following lemma shows that unique $\hat{\boldsymbol{\beta}}_L$ exists when $p_n \leq n$. This lemma holds for logistic, probit and quadratic loss functions, as they are strictly convex and non-negative.

Lemma A.24. *If* $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \in \mathbb{R}^{p_n} \times \{0, 1\}$ *is a random sample,* $\rho(\cdot, y)$ *is strictly convex function bounded from below by* $m \in \mathbb{R}$, $p_n \leq n$, $\lambda > 0, \mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ *and* $\operatorname{rank}\mathbb{X} = p_n$, *then exists unique*

$$\hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{b}\in\mathbb{R}^{p_n}}{\arg\min} \left(\frac{1}{n}\sum_{i=1}^{n} \rho(\boldsymbol{b}^T\boldsymbol{X}_i, Y_i) + \lambda||\boldsymbol{b}||_1 \right).$$

Proof. We note that for $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^{p_n}$ with $\mathbf{b}_1 \neq \mathbf{b}_2$ we have $\mathbb{X}\mathbf{b}_1 \neq \mathbb{X}\mathbf{b}_2$, as $\operatorname{rank}\mathbb{X} = p_n$. Hence there exists $i_0 \in \{1, \ldots, n\}$ such that $\mathbf{b}_1^T\mathbf{X}_i \neq \mathbf{b}_2^T\mathbf{X}_i$. Let $h_i(\mathbf{b}) = \rho(\mathbf{b}^T\mathbf{X}_i, Y_i)$ for $i = 1, \ldots, n$. Strict convexity of ρ gives for $\alpha \in [0, 1]$:

$$h_{i_0}(\alpha\mathbf{b}_1 + (1-\alpha)\mathbf{b}_2) < \alpha h_{i_0}(\mathbf{b}_1) + (1-\alpha)h_{i_0}(\mathbf{b}_2).$$

Hence h_{i_0} is strictly convex. Moreover h_i are convex from convexity of ρ for all $i \in \{1, \ldots, n\}$. Function:

$$P_n(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{b}^T\mathbf{X}_i, Y_i) + \lambda||\mathbf{b}||_1 = \frac{1}{n}\sum_{i=1}^{n} h_i(\mathbf{b}) + \lambda||\mathbf{b}||_1$$

is strictly convex as a sum of strictly convex and convex functions. This means that P_n is continuous. Moreover, $P_n(\mathbf{b})$ is coercive, as we have:

$$P_n(\mathbf{b}) \geq m + \lambda||\mathbf{b}||_1 \to +\infty$$

for $||\mathbf{b}||_1 \to +\infty$. Thus the existence of $\hat{\boldsymbol{\beta}}_L$ follows from Lemma A.3. Uniqueness of $\hat{\boldsymbol{\beta}}_L$ follows from strict convexity of P_n. □

Proofs of Lemmas A.25, A.26 and Theorem A.27 below are strengthtened versions of the proof of Theorem 3 in Rosset et al. (2004) which do not use differentiability of loss function. Theorem A.27 shows that there exist sparse Lasso solutions having at most n nonzero coefficients, when $p_n \geq n$ for a general loss function ρ provided a solution exists.

Lemma A.25. *Let* $s\colon\ \mathbb{R} \to \mathbb{R}$ *be defined as* $s(t) = ||\boldsymbol{b} + t\boldsymbol{a}||_1$ *for* $t \in \mathbb{R}$, $\boldsymbol{b}, \boldsymbol{a} \in \mathbb{R}^{p_n}$ *and* $\boldsymbol{a} \neq \boldsymbol{0}_{p_n}$. *Then there exists* $t^* = \arg\min s(t)$ *and* $i \in \operatorname{supp} \boldsymbol{a}$ *satisfying* $b_i + t^* a_i = 0$.

Proof. Let $S = \{k \in \{1, \ldots, p_n\}\colon\ a_k \neq 0\}$. Then we obtain:

$$s(t) = \sum_{k \in S^c} |b_k| + \sum_{k \in S} |a_k| \cdot \left| \frac{b_k}{a_k} + t \right|.$$

Without losing of generality we can assume that $S = \{k_1, \ldots, k_l\}$ for some $l \in \mathbb{N}$ and:

$$\frac{b_{k_1}}{a_{k_1}} \geq \ldots \geq \frac{b_{k_l}}{a_{k_l}}.$$

Let

$$m = \min \left\{ s\left(-\frac{b_{k_1}}{a_{k_1}}\right), \ldots, s\left(-\frac{b_{k_l}}{a_{k_l}}\right) \right\}.$$

Since for every $i \in \{2, \ldots, l\}$ and $t \in \left[-b_{k_{i-1}}/a_{k_{i-1}}, -b_{k_i}/a_{k_i}\right]$ function s is linear, we obtain for such t

$$s(t) \geq \min \left\{ s\left(-\frac{b_{k_{i-1}}}{a_{k_{i-1}}}\right), s\left(-\frac{b_{k_i}}{a_{k_i}}\right) \right\} \geq m.$$

Analogously, from the fact that $\lim_{|t| \to \pm\infty} s(t) = +\infty$, for $t \in I_1 = (-\infty, -b_{k_1}/a_{k_1}]$ and $t \in I_2 = [-b_{k_l}/a_{k_l}, +\infty)$ and from linearity of s we obtain

$$s(t) \geq s\left(-\frac{b_{k_1}}{a_{k_1}}\right) \geq m \text{ and } s(t) \geq s\left(-\frac{b_{k_l}}{a_{k_l}}\right) \geq m$$

for t belonging to I_1 and I_2 respectively. We thus have: $s(t) \geq m$ for $t \in \mathbb{R}$. Hence for some $t^* = -b_i/a_i$ function s achieves its minimum, what proves our claim. □

Lemma A.26. *Let* $g\colon\ \mathbb{R}^n \to \mathbb{R}$ *be some function and let* $h\colon\ \mathbb{R}_+ \cup \{0\} \to \mathbb{R}$ *be non-decreasing function. Let* $\mathbb{X} \in \mathbb{R}^{n \times p_n}$ *and* $f\colon\ \mathbb{R}^{p_n} \to \mathbb{R}$ *- a function defined as:*

$$f(\boldsymbol{b}) = g(\mathbb{X}\boldsymbol{b}) + h(||\boldsymbol{b}||_1).$$

Then for every $\boldsymbol{b} \in \mathbb{R}^{p_n}$ *such that* $|\operatorname{supp} \boldsymbol{b}| > \operatorname{rank} \mathbb{X}$ *there exists* $\boldsymbol{c} \in \mathbb{R}^{p_n}$ *such that* $|\operatorname{supp} \boldsymbol{c}| \leq \operatorname{rank} \mathbb{X}$ *and* $f(\boldsymbol{c}) \leq f(\boldsymbol{b})$.

Proof. Columns of matrix $\mathbb{X}_{\operatorname{supp} \mathbf{b}}$ are linearly dependent as $|\operatorname{supp} \mathbf{b}| > \operatorname{rank} \mathbb{X}$. This means that exists $\mathbf{a} \in \mathbb{R}^{p_n} \setminus \{\mathbf{0}_{p_n}\}$ such that $\operatorname{supp} \mathbf{a} \subseteq \operatorname{supp} \mathbf{b}$ and $\mathbb{X}\mathbf{a} = 0$. Now we consider a function $d\colon\ \mathbb{R} \to \mathbb{R}$:

$$d(t) = f(\mathbf{b} + t\mathbf{a}) = g(\mathbb{X}(\mathbf{b} + t\mathbf{a})) + h(||\mathbf{b} + t\mathbf{a}||_1) = g(\mathbb{X}\mathbf{b}) + h(||\mathbf{b} + t\mathbf{a}||_1).$$

To find the minimum of function d, we have to minimize $s(t) = ||\mathbf{b} + t\mathbf{a}||_1$ for $t \in \mathbb{R}$, as h is non-decreasing. Function s is convex, therefore its minimum exists. In view of Lemma A.25 there exists $t^* = \arg\min s(t)$ and exists $i \in \operatorname{supp} b$ such that $b_i + t^* a_i = 0$. We take that t^* and define $\mathbf{c}_1 = \mathbf{b} + t^*\mathbf{a}$. Then $d(t^*) \leq d(0)$, what implies $f(\mathbf{c}_1) \leq f(\mathbf{b})$. From our choice of t^* we get $\operatorname{supp} \mathbf{c}_1 \subset \operatorname{supp} \mathbf{b}$ and $|\operatorname{supp} \mathbf{c}_1| \leq |\operatorname{supp} \mathbf{b}| - 1$. If $|\operatorname{supp} \mathbf{c}_1| \leq \operatorname{rank} \mathbb{X}$, we take $\mathbf{c} = \mathbf{c}_1$ and the lemma is proven. If not, then we iterate this procedure (by setting $\mathbf{b} := \mathbf{c}_1$) and after finite number of steps we obtain $\mathbf{c}$ having the desired properties. □

Theorem A.27. *Let* $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \in \mathbb{R}^{p_n} \times \{0, 1\}$ *is a random sample,* $\rho\colon\ \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ *and* $\lambda > 0$. *Let*

$$P_n(\boldsymbol{b}) = \frac{1}{n} \sum_{i=1}^{n} \rho(\boldsymbol{b}^T \boldsymbol{X}_i, Y_i) + \lambda ||\boldsymbol{b}||_1$$

and assume that there exists

$$\boldsymbol{c} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} P_n(\boldsymbol{b}).$$

Then there exists $\boldsymbol{b}_0 = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} P_n(\boldsymbol{b})$ *such that* $|\operatorname{supp} \boldsymbol{b}_0| \leq \operatorname{rank} \mathbb{X} \leq n$.

Proof. Our proof starts with the observation that if $|\operatorname{supp} \mathbf{c}| \leq \operatorname{rank} \mathbb{X}$, then we take $\mathbf{b}_0 = \mathbf{c}$. If not, then in view of Lemma A.26 there exists $\mathbf{a}$ such that $P_n(\mathbf{a}) \leq P_n(\mathbf{c})$ and $|\operatorname{supp} \mathbf{a}| \leq \operatorname{rank} \mathbb{X}$. This means that $\mathbf{a} = \arg\min_{\mathbf{b} \in \mathbb{R}^p} P_n(\mathbf{b})$ and we take $\mathbf{b}_0 = \mathbf{a}$. □

Theorem A.28. *Let* $\rho(\cdot, y)$ *be a differentiable function for all* y. *Assume that there does not exist set* J, $\boldsymbol{a} \in \mathbb{R}^{|J|} \setminus \{\boldsymbol{0}_{|J|}\}$ *and* $\boldsymbol{\sigma} \in \{-1, 1\}^{|J|}$ *such that* $|J| > n$, $\mathbb{X}_J \boldsymbol{a} = \boldsymbol{0}_n$ *and* $\boldsymbol{\sigma}^T \boldsymbol{a} = 0$. *Then every vector* $\hat{\boldsymbol{\beta}}_L$ *minimizing* P_n *defined in Theorem A.27 has at most* n *nonzero coefficients.*

Proof. Suppose the assertion of the theorem is false. Then exists $\hat{\boldsymbol{\beta}}_L$ minimizing P_n which has more than n nonzero coefficients. Let $J = \{j \in \{1, \ldots, p_n\}\colon\ \hat{\boldsymbol{\beta}}_{L,j} \neq 0\}$. Equation (4.10) for indices in J takes the form:

$$\lambda \operatorname{sgn} \hat{\boldsymbol{\beta}}_{L,J}^T = \mathbf{v}^T \mathbb{X}_J.$$

By our assumption about $\hat{\boldsymbol{\beta}}_L$, we have $|J| > n$ and there exists $\mathbf{a} \in \mathbb{R}^{|J|} \setminus \{\mathbf{0}_{|J|}\}$ such that $\mathbb{X}_J \mathbf{a} = \mathbf{0}_{|J|}$. Thus we get: $\lambda \operatorname{sgn} \hat{\boldsymbol{\beta}}_{L,J}^T \mathbf{a} = \mathbf{v}^T \mathbb{X}_J \mathbf{a} = 0$. Taking $\boldsymbol{\sigma} = \operatorname{sgn} \hat{\boldsymbol{\beta}}_{L,J} \in \{-1, 1\}^{|J|}$ proves the theorem by contradiction. □

For completeness we state two known results which concern uniqueness of the solution defined in Theorem A.27.

Theorem A.29. *If assumptions of Theorem A.28 are satisfied,* $\rho(\cdot, y)$ *is strictly convex for all* y *and for every* $M \subset \{1, \ldots, p_n\}$ *with* $|M| \leq n$ *columns of* $\mathbb{X}_M$ *are linearly independent, then* $\hat{\boldsymbol{\beta}}_L$ *minimizing* P_n *defined in Theorem A.27 is unique.*

Proof. Proof is identical with the proof of Theorem 5 in Rosset et al. (2004). □

Theorem A.30. *(Lemma 5 in Tibshirani (2013)) If $\mathbb{X} \in \mathbb{R}^{n \times p_n}$ has entries drawn from a continuous probability distribution on $\mathbb{R}^{np_n}$, $\rho(\cdot, y)$ is differentiable, strictly convex function for all y and $\rho(b, y) > -\infty$ for all b, y, then for any $\lambda > 0$ $\hat{\boldsymbol{\beta}}_L$ minimizing P_n defined in Theorem A.27 is unique with probability 1 and this solution has at most $\min\{n, p_n\}$ nonzero coefficients.*

Note that result of theorem above holds in particular for quadratic and logistic loss and also for dependent observations.

A.4. Selected properties of subgaussian random variables

In this section we present definition and basic properties of subgaussian random variables which are used in Chapters 4-5.

Definition A.31. *We call a random variable $X \in \mathbb{R}$ subgaussian if there exists $\sigma \geq 0$ that for all $t \in \mathbb{R}$ we have $\mathbb{E}\exp(tX) \leq \exp(t^2\sigma^2/2)$. If variable X satisfies this property, we will write $X \sim Subg(\sigma^2)$.*

Lemma A.32. *If $X \sim Subg(\sigma^2)$, then we have:*

1. $\mathbb{E}X = 0$,
2. $\mathbb{E}X^2 \leq \sigma^2$,
3. *for all $t \geq 0$:* $\mathbb{P}(|X| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$,
4. *for all $p \geq 1$:* $\mathbb{E}|X|^p \leq p\sigma^p\sqrt{2}^p\Gamma\left(\frac{p}{2}\right)$,
5. *for all $\eta \in [0, 1)$:* $\mathbb{E}\exp\left(\frac{\eta X^2}{2\sigma^2}\right) \leq \frac{1}{\sqrt{1-\eta}}$.

Proof. For the proofs of statements 1 and 2 we use inequality:

$$1 + t\mathbb{E}X + \frac{t^2\mathbb{E}X^2}{2} \leq \mathbb{E}e^{tX} \leq e^{\frac{t^2\sigma^2}{2}}.$$

Using $\lim_{x\to 0}(e^x - 1)/x = 1$ and above inequality yields:

$$\mathbb{E}X \leq \lim_{t\to 0^+}\left(\frac{e^{\frac{t^2\sigma^2}{2}} - 1}{\frac{t^2\sigma^2}{2}} \cdot \frac{t\sigma^2}{2} - \frac{t\mathbb{E}X^2}{2}\right) = 0$$

and

$$\mathbb{E}X \geq \lim_{t\to 0^-}\left(\frac{e^{\frac{t^2\sigma^2}{2}} - 1}{\frac{t^2\sigma^2}{2}} \cdot \frac{t\sigma^2}{2} - \frac{t\mathbb{E}X^2}{2}\right) = 0.$$

This means that $\mathbb{E}X = 0$. Proof of statement 2 can be conducted in the same fashion, using $\mathbb{E}X = 0$:

$$\mathbb{E}X^2 \leq \lim_{t\to 0}\frac{e^{\frac{t^2\sigma^2}{2}} - 1}{\frac{t^2\sigma^2}{2}} \cdot \sigma^2 = \sigma^2.$$

To prove statement 3, firstly we use Chernoff's inequality for $\lambda > 0$ and inequality $\max\{a, b\} \leq a + b$ for $a, b \geq 0$:

$$\mathbb{P}(|X| \geq t) \leq e^{-\lambda t}\mathbb{E}e^{\lambda|X|} = e^{-\lambda t}\mathbb{E}\max\{e^{\lambda X}, e^{-\lambda X}\} \leq e^{-\lambda t}(\mathbb{E}e^{\lambda X} + \mathbb{E}e^{-\lambda X}) \leq 2e^{-\lambda t + \frac{\lambda^2\sigma^2}{2}}.$$

Taking optimal $\lambda = t/\sigma^2$ gives statement 3. Proof of statement 4 uses known representation of moments of random variables, statement 3 and moments of normal distribution:

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}(|X| \geq t)pt^{p-1}dt \leq \int_0^\infty 2e^{-\frac{t^2}{2\sigma^2}}pt^{p-1}dt = p\sigma^p\sqrt{2}^p\Gamma\left(\frac{p}{2}\right).$$

Proof of statement 5 can be found in Lemma 7.4 in Baraniuk et al. (2011). □

Remark A.33. *If random variable X is bounded: $X \in [a, b]$ for some $a \leq b$ and $\mathbb{E}X = 0$, then $X \sim Subg((b-a)^2/4)$.*

Proof. See Lemma 2.6 in Massart (2007) for proof. □

The following two lemmas show that sum of subgaussian random variables is always subgaussian. Moreover, in the case of dependent random variables, Lemma A.35 gives worse subgaussianity constant than Lemma A.34 and thus results from Chapters 4-5 cannot be easily generalized to the case of dependent observations.

Lemma A.34. *If $X_i \sim Subg(\sigma_i^2)$ for $i = 1, \ldots, n$ are independent then $\sum_{i=1}^n X_i \sim Subg\left(\sum_{i=1}^n \sigma_i^2\right)$.*

Proof. From independence and subgaussianity of X_i for $i = 1, \ldots, n$ we have:

$$\mathbb{E}e^{t\sum_{i=1}^n X_i} = \prod_{i=1}^n \mathbb{E}e^{tX_i} \leq \prod_{i=1}^n e^{\frac{t^2\sigma_i^2}{2}} = e^{\frac{t^2\sum_{i=1}^n \sigma_i^2}{2}}.$$

□

Lemma A.35. *If $X_i \sim Subg(\sigma_i^2)$ then $\sum_{i=1}^n X_i \sim Subg\left(\left(\sum_{i=1}^n \sigma_i\right)^2\right)$.*

Proof. We prove this lemma by induction. If $n = 1$, then it is obvious. Assume that lemma is true for some $n \in \mathbb{N}_+$. Then for $n + 1$ in view of Hölder's inequality, subgaussianity of X_{n+1} and induction assumption we have for $\lambda_1, \lambda_2 \geq 1$ such that $\lambda_1^{-1} + \lambda_2^{-1} = 1$:

$$\mathbb{E}e^{t\sum_{i=1}^{n+1} X_i} \leq \left(\mathbb{E}e^{t\lambda_1 X_{n+1}}\right)^{\frac{1}{\lambda_1}}\left(\mathbb{E}e^{t\lambda_2\sum_{i=1}^n X_i}\right)^{\frac{1}{\lambda_2}} \leq \exp\left(\frac{t^2\sigma_{n+1}^2\lambda_1}{2}\right)\exp\left(\frac{t^2\lambda_2\left(\sum\limits_{i=1}^n \sigma_i\right)^2}{2}\right). \tag{A.6}$$

We take

$$\lambda_1 = \frac{\sum\limits_{i=1}^{n+1}\sigma_i}{\sigma_{n+1}}, \quad \lambda_2 = \frac{\sum\limits_{i=1}^{n+1}\sigma_i}{\sum\limits_{i=1}^n \sigma_i}.$$

Note that

$$\frac{1}{\lambda_1} + \frac{1}{\lambda_2} = \frac{\sigma_{n+1}}{\sum\limits_{i=1}^{n+1}\sigma_i} + \frac{\sum\limits_{i=1}^n \sigma_i}{\sum\limits_{i=1}^{n+1}\sigma_i} = 1.$$

Thus (A.6) gives:

$$\mathbb{E}e^{t\sum_{i=1}^{n+1}X_i} \leq \exp\left(\frac{t^2\sigma_{n+1}^2\lambda_1}{2}\right)\exp\left(\frac{t^2\lambda_2\left(\sum_{i=1}^{n}\sigma_i\right)^2}{2}\right)$$

$$= \exp\left(\frac{t^2\sigma_{n+1}\left(\sum_{i=1}^{n+1}\sigma_i\right) + t^2\left(\sum_{i=1}^{n}\sigma_i\right)\cdot\left(\sum_{i=1}^{n+1}\sigma_i\right)}{2}\right) = \exp\left(\frac{t^2\left(\sum_{i=1}^{n+1}\sigma_i\right)^2}{2}\right).$$

This ends the proof. □

Lemma A.36. *Let $Z_j \sim Subg(\sigma^2)$ for $j \in \mathcal{A}$ and $|\mathcal{A}| > 1$. Then*

$$\mathbb{E}\max_{j\in\mathcal{A}}|Z_j| \leq \frac{7}{2}\sigma\sqrt{\ln|\mathcal{A}|} \tag{A.7}$$

Proof. Using union inequality and statement 3 from Lemma A.32 gives

$$\mathbb{P}(\max_{j\in\mathcal{A}}|Z_j| > t) \leq \sum_{j\in\mathcal{A}}\mathbb{P}(|Z_j| > t) \leq 2|\mathcal{A}|\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Hence for any positive c:

$$\mathbb{E}\max_{j\in\mathcal{A}}|Z_j| = \int_0^\infty \mathbb{P}\left(\max_{j\in\mathcal{A}}|Z_j| > t\right)dt \leq c + 2|\mathcal{A}|\int_c^\infty \exp\left(-\frac{t^2}{2\sigma^2}\right)dt$$

$$\leq c + \frac{2|\mathcal{A}|}{c}\int_c^\infty t\exp\left(-\frac{t^2}{2\sigma^2}\right)dt = c + \frac{2|\mathcal{A}|}{c}\sigma^2\exp\left(-\frac{c^2}{2\sigma^2}\right).$$

For $c = \sqrt{2\sigma^2\ln|\mathcal{A}|}$ we obtain:

$$\mathbb{E}\max_{j\in\mathcal{A}}|Z_j| \leq \sqrt{2\sigma^2}\left(\sqrt{\ln|\mathcal{A}|} + \frac{1}{\sqrt{\ln|\mathcal{A}|}}\right) \leq \frac{7}{2}\sigma\sqrt{\ln|\mathcal{A}|},$$

where the last inequality uses the fact that

$$\frac{\sqrt{2}}{\sqrt{\ln|\mathcal{A}|}} \leq \frac{\sqrt{2}}{\sqrt{\ln 2}} \leq 2.05\sqrt{\ln 2} \leq 2.05\sqrt{\ln|\mathcal{A}|}$$

and that $2.05 + \sqrt{2} \leq 7/2$. □

Below we give an auxiliary proof of known inequality for Γ function, which will be used in Lemma A.38. It is strengthened version of inequality in Lemma 1 in Minc and Sathre (1964):

$$\log(\Gamma(x)) - ((x-1/2)\log(x) - x + \log(2\pi)/2) < 1/x < 1 \quad \forall\ x > 1$$

and it uses ideas from the proof of that Lemma.

Lemma A.37.

$$\Gamma(x) < \sqrt{2\pi}x^{x-\frac{1}{2}}e^{-x+\frac{1}{12x}}$$

for $x > 0$.

Proof. We use the following Binet's formula (see Bateman (1953, Chapter I, Section 1.9 Equation (4), p.21)):

$$\log\Gamma(z) = \left(z-\frac{1}{2}\right)\log z - z + \frac{1}{2}\log(2\pi) + \int_0^\infty \left(\frac{1}{e^t-1} - \frac{1}{t} + \frac{1}{2}\right)\cdot\frac{1}{t}e^{-tz}dt, \tag{A.8}$$

where $\operatorname{Re} z > 0$. Now we prove that for $t > 0$:

$$\left(\frac{1}{e^t-1} - \frac{1}{t} + \frac{1}{2}\right)\cdot\frac{1}{t} \leq \frac{1}{12}.$$

Multiplying both sides by $12t^2(e^t-1)$ and arranging terms shows that we need to prove:

$$(e^t-1)(t^2-6t+12) > 12t. \tag{A.9}$$

Observe that

$$e^t \geq 1 + t + \frac{t^2}{2} + \frac{t^3}{6} + \frac{t^4}{24} + \frac{t^5}{120}.$$

Thus using the above inequality and $t^2 - t + 2 > t^2 - 2t + 2 = (t-1)^2 + 1 > 0$ gives for $t > 0$:

$$(e^t-1)(t^2-6t+12) \geq \left(t + \frac{t^2}{2} + \frac{t^3}{6} + \frac{t^4}{24} + \frac{t^5}{120}\right)\cdot(t^2-6t+12) \tag{A.10}$$

$$= \frac{1}{120}t^5(t^2-t+2) + 12t > 12t. \tag{A.11}$$

This means that for $x > 0$:

$$\log\Gamma(x) < \left(x-\frac{1}{2}\right)\log x - x + \frac{1}{2}\log(2\pi) + \frac{1}{12}\int_0^\infty e^{-tx}dt$$

$$= \left(x-\frac{1}{2}\right)\log x - x + \frac{1}{2}\log(2\pi) + \frac{1}{12x},$$

what proves our statement. □

We state first two auxiliary Lemmas which will be used in proofs of Lemmas 4.8, 4.14 and 5.1. The following Lemma used in the proof of Lemma 4.8 which is interesting in its own right states that a product of a subgaussian random variable by a bounded one is subgaussian provided it has expectation zero. Explicit value of subgaussianity parameter is provided.

Lemma A.38. *Assume that $S \sim Subg(\sigma^2)$ and T is random variable such that $|T| \leq M$, where M is some positive constant and $\mathbb{E}(ST) = 0$. Then $ST \sim Subg(\tau^2M^2\sigma^2)$, where $\tau = e^{\frac{13}{24}}\cdot 4/\sqrt[4]{27} \leq 3.02$.*

Proof. Since S is subgaussian and T is bounded, we obtain using Lemma A.32 p. 3:

$$\mathbb{P}(|ST| \geq t) \leq \mathbb{P}\left(|S| \geq \frac{t}{M}\right) \leq 2e^{\frac{-t^2}{2\sigma^2M^2}}. \tag{A.12}$$

From the above inequality, by using the same argument as in Vershynin (2012), it follows that for $p \geq 2$:

$$\mathbb{E}|ST|^p = \int_0^\infty \mathbb{P}(|ST| \geq t)pt^{p-1}dt \leq \int_0^\infty 2e^{-\frac{t^2}{2M^2\sigma^2}}pt^{p-1}dt = pM^p\sigma^p\sqrt{2}^p\Gamma\left(\frac{p}{2}\right). \tag{A.13}$$

By applying the above inequality and well known inequalities (see Robbins (1955) and Lemma A.37):

$$\Gamma\left(\frac{p}{2}\right) \leq \sqrt{2\pi}\left(\frac{p}{2}\right)^{\frac{p}{2}-\frac{1}{2}} e^{-\frac{p}{2}+\frac{1}{6p}}, \qquad p! \geq \sqrt{2\pi}p^{p+\frac{1}{2}}e^{-p},$$

we obtain using $\mathbb{E}ST = 0$

$$\begin{aligned}
\mathbb{E}e^{tST} &= 1 + \sum_{p=2}^{\infty} \frac{t^p \mathbb{E}(ST)^p}{p!} \leq 1 + \sum_{p=2}^{\infty} \frac{|t|^p M^p \sigma^p 2\sqrt{\pi} p^{\frac{p}{2}+\frac{1}{2}} e^{-\frac{p}{2}+\frac{1}{6p}}}{p!} \\
&\leq 1 + \sum_{p=2}^{\infty} \frac{|t|^p M^p \sigma^p 2\sqrt{\pi} p^{\frac{p}{2}+\frac{1}{2}} e^{-\frac{p}{2}+\frac{1}{6p}}}{\sqrt{2\pi}p^{p+\frac{1}{2}}e^{-p}} = 1 + \sum_{p=2}^{\infty} \left(\frac{|t|M\sigma\sqrt{e}}{\sqrt{p}}\right)^p \sqrt{2} e^{\frac{1}{6p}} \\
&\leq 1 + \sum_{p=2}^{\infty} \left(\frac{|t|M\sigma\sqrt{e}}{\sqrt{p}}\right)^p \sqrt{2} e^{\frac{1}{12}}. \qquad \text{(A.14)}
\end{aligned}$$

Observe that for $k \geq 2$ we have (see Robbins (1955)):

$$k! < \sqrt{2\pi}k^{k+\frac{1}{2}}e^{-k+\frac{1}{12k}} \leq \sqrt{2\pi}e^{\frac{1}{24}}k^{k+\frac{1}{2}}e^{-k} \leq ek^{k+\frac{1}{2}}e^{-k}.$$

Hence for $k \geq 1$ (for $k = 1$ both sides of first inequality are equal):

$$k! \leq ek^{k+\frac{1}{2}}e^{-k} \leq e^{\frac{1}{2}}k^k e^{-\frac{k}{2}}.$$

Thus we obtain for $C \geq 0$:

$$e^{C^2t^2} = 1 + \sum_{k=1}^{\infty} \frac{(t^2C^2)^k}{k!} \geq 1 + \sum_{k=1}^{\infty} \left(\frac{C|t|e^{\frac{1}{4}}}{\sqrt{k}}\right)^{2k} \cdot e^{-\frac{1}{2}}. \qquad \text{(A.15)}$$

In order to show that $\mathbb{E}e^{tST} \leq e^{C^2t^2}$, we prove that the series in (A.15) bounds from above the sum appearing in the bound of $\mathbb{E}e^{tST}$. To this end, consider the function

$$f(x) = \frac{x^x(x+1)^{x+1}}{(x+\frac{1}{2})^{2x+1}}$$

which is decreasing for $x \geq 1$ and thus

$$f(x) \leq f(1) = \frac{32}{27}.$$

This implies

$$k^k(k+1)^{(k+1)} \leq \frac{32}{27}\left(k+\frac{1}{2}\right)^{2k+1}.$$

Hence from the inequality $x^2 + y^2 \geq 2xy$ and the inequality above we have

$$\left(\frac{C|t|e^{\frac{1}{4}}}{\sqrt{k}}\right)^{2k} + \left(\frac{C|t|e^{\frac{1}{4}}}{\sqrt{k+1}}\right)^{2k+2} \geq \frac{2(C|t|e^{\frac{1}{4}})^{2k+1}}{\sqrt{k}^k\sqrt{k+1}^{k+1}} \geq \sqrt{\frac{27}{8}}\left(\frac{C|t|e^{\frac{1}{4}}}{\sqrt{k+\frac{1}{2}}}\right)^{2k+1}. \qquad \text{(A.16)}$$

Define for $a \geq 0$

$$I_a = \sum_{k=1}^{\infty} \left(\frac{C|t|e^{\frac{1}{4}}}{\sqrt{k+a}}\right)^{2(k+a)} \text{ and } I = \sum_{p=2}^{\infty} \left(\frac{C|t|e^{\frac{1}{4}}\sqrt{2}}{\sqrt{p}}\right)^p.$$

From inequalities $I_0 \geq I_1$, (A.16), $I_0 \geq 0$ and $I_0 + I_{\frac{1}{2}} = I$ we have

$$I_0 \geq \frac{3}{4}I_0 + \frac{1}{4}I_1 \geq \frac{1}{2}I_0 + \sqrt{\frac{27}{128}}I_{\frac{1}{2}} \geq \sqrt{\frac{27}{128}}I.$$

From the last inequality and (A.15) we obtain

$$e^{C^2t^2} \geq 1 + e^{-\frac{1}{2}}I_0 \geq 1 + e^{-\frac{1}{2}}\sqrt{\frac{27}{128}}I. \qquad \text{(A.17)}$$

Note that for $C \geq M\sigma e^{\frac{13}{24}}\sqrt[4]{64/27}$ we have

$$\sum_{p=2}^{\infty}\left(\frac{|t|M\sigma\sqrt{e}}{\sqrt{p}}\right)^p\sqrt{\frac{256}{27}}e^{\frac{7}{12}} \leq \sum_{p=2}^{\infty}\left(\frac{|t|M\sigma e^{\frac{1}{2}}}{\sqrt{p}}\cdot e^{\frac{7}{24}}\cdot\sqrt[4]{\frac{256}{27}}\right)^p \leq I. \tag{A.18}$$

From (A.18) and the bound for $\mathbb{E}e^{tST}$ in (A.14) we have

$$\mathbb{E}e^{tST} \leq 1 + e^{1/12}\sqrt{\frac{27}{256}}e^{-7/12}\times I \leq e^{C^2t^2} \tag{A.19}$$

for $C \geq M\sigma e^{\frac{13}{24}}\sqrt[4]{64/27}$, where the last inequality in (A.19) follows from (A.17). This ends the proof. □

The following Lemma is a version of Lemma A.38 for independent variables S and T and is used in the proof of Lemmas 4.14 and 5.1. Note that it gives smaller subgaussianity constant than Lemma A.38.

Lemma A.39. *Assume that $S \sim Subg(\sigma^2)$ and T is random variable such that $|T| \leq M$, where M is some positive constant and S and T are independent. Then $ST \sim Subg(M^2\sigma^2)$.*

Proof. Observe that:

$$\mathbb{E}e^{tST} = \mathbb{E}(\mathbb{E}(e^{tST}|T)) \leq \mathbb{E}e^{\frac{t^2T^2\sigma^2}{2}} \leq e^{\frac{t^2M^2\sigma^2}{2}}.$$

□

A.5. Inequalities related to Rademacher averages

Theorems in this section are useful for finding expectation bounds for expressions of the form:

$$\sup_{\mathbf{b}\in A:||\mathbf{b}-\boldsymbol{\beta}^*||_p\leq r}|(R_n(\mathbf{b}) - R(\mathbf{b})) - (R_n(\boldsymbol{\beta}^*) - R(\boldsymbol{\beta}^*))|, \tag{A.20}$$

where empirical risk

$$R_n(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}\rho(\mathbf{b}^T\mathbf{X}_i, Y_i)$$

was defined in 4.3, $R(\mathbf{b}) = \mathbb{E}R_n(\mathbf{b})$, $A \subseteq \mathbb{R}^p$ and $p \geq 1$. Note that the symmetrization inequality given below in Theorem A.40 is a special case of Lemma 2.3.1 in van der Vaart and Wellner (1996). The version given below is sufficient in our applications. Theorem A.41 was originally formulated in Ledoux and Talagrand (1991) for contractions, but we observe that it holds for Lipschitz function $g : \mathbb{R} \to \mathbb{R}$ with constant $L > 0$, if we take g/L instead of g. Moreover, assumption $g(0) = 0$ in can be easily omitted by taking $\rho(b, y) - \rho(0, y)$ instead of $\rho(b, y)$ in Lemmas 4.14 and 5.1. Boundedness of $f(\mathbf{X}_i)$ assumed in Theorem 4.12 in Ledoux and Talagrand (1991) is not needed in the Theorem A.41, because we can prove the result conditionally on $(\mathbf{X}_i)_i$, assume integrability of appropriate functions and take expectations of both sides.

Theorem A.40 (Symmetrization inequality, see van der Vaart and Wellner (1996, Lemma 2.3.1))**.** *Let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be nondecreasing, convex function and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent random variables with values in $\mathbb{R}^p$. Let $\mathcal{F}$ be some set of measurable functions $f\colon \mathbb{R}^p \to \mathbb{R}$. Assume that: $\mathbb{E}\Phi(2\sup_{f\in\mathcal{F}}|f(\boldsymbol{X}_i)|) < \infty$, $\sup_{f\in\mathcal{F}}\mathbb{E}|f(\boldsymbol{X}_i)| < \infty$ for all $i = 1, \ldots, n$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher variables independent of $(\boldsymbol{X}_i)_{i=1,\ldots,n}$, ie. $\mathbb{P}(\varepsilon_i = \pm 1) = 0.5$. Then we have:*

$$\mathbb{E}\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{X}_i) - \mathbb{E}f(\boldsymbol{X}_i))\right|\right) \leq \mathbb{E}\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(\boldsymbol{X}_i)\right|\right).$$

Theorem A.41 (Talagrand-Ledoux inequality, see Ledoux and Talagrand (1991, Theorem 4.12))**.** *Let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be convex and increasing and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent random variables with values in $\mathbb{R}^p$. Let $\mathcal{F}$ be some set of measurable functions $f\colon \mathbb{R}^p \to \mathbb{R}$. Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz function with constant $L > 0$ and $g(0) = 0$. Assume that $\mathbb{E}\Phi(2L\sup_{f\in\mathcal{F}}|f(X_i)|) < \infty$ for all i. Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher variables independent of $(\boldsymbol{X}_i)_{i=1,\ldots,n}$. Then we have:*

$$\mathbb{E}\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(f(\boldsymbol{X}_i))\right|\right) \leq \mathbb{E}\Phi\left(2L\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(\boldsymbol{X}_i)\right|\right).$$

A.6. Lasso consistency for logistic regression with intercept

We consider setup of Chapter 4 for model with intercept when logistic lasso is fitted.

The following theorem is a modification of Theorem 5 in Fan et al. (2014a) which was stated for logistic model without intercept. The proof is based mainly on the proof of that theorem (see Fan et al. (2014b)) and only differences in key inequalities are written down. The crucial difference in the present proof is a term $|\hat{\boldsymbol{\beta}}_{L,0} - \boldsymbol{\beta}_0^*|$ in (A.21) (compare (2) in Fan et al. (2014b)).

Theorem A.42. *If $(\boldsymbol{X}_i, Y_i)_{i=1,\ldots,n}$ are random variables such that $||DR_n(\boldsymbol{\beta}^*)||_\infty \leq \frac{\lambda}{2}$, $\lambda \leq \frac{\kappa_n}{20K|s_0^*|}$, where*

$$\kappa_n = \inf_{\boldsymbol{\Delta}\in\mathcal{C}} \frac{\boldsymbol{\Delta}^T D^2 R_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\boldsymbol{\Delta}},$$

$$\mathcal{C} = \{\boldsymbol{\Delta} \in \mathbb{R}^{p_n+1} : 3||\boldsymbol{\Delta}_{s_0^*}||_1 \geq ||\tilde{\boldsymbol{\Delta}}_{s^{*c}}||_1\},$$

$$s_0^* = s^* \cup \{0\}$$

and

$$K = \max_{\substack{i=1,\ldots,n\\ j=0,1,\ldots,p_n}} |X_{ij}|,$$

then we have:

$$||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_2 \leq 5|s_0^*|^{1/2}\lambda\kappa_n^{-1}.$$

Proof. By definition of $\hat{\boldsymbol{\beta}}_L$ we have:

$$R_n(\hat{\boldsymbol{\beta}}_L) + \lambda||\hat{\tilde{\boldsymbol{\beta}}}_L||_1 \leq R_n(\boldsymbol{\beta}^*) + \lambda||\tilde{\boldsymbol{\beta}}^*||_1.$$

Definition of $\hat{\boldsymbol{\beta}}_L$, convexity of $R_n(\boldsymbol{\beta}^*)$ and condition $||DR_n(\boldsymbol{\beta}^*)||_\infty \leq \frac{\lambda}{2}$ yield:

$$\begin{aligned}||\tilde{\boldsymbol{\beta}}^*||_1 &\geq \lambda^{-1}(R_n(\hat{\boldsymbol{\beta}}_L) - R_n(\boldsymbol{\beta}^*)) + ||\hat{\tilde{\boldsymbol{\beta}}}_L||_1 \geq \lambda^{-1}DR_n(\boldsymbol{\beta}^*)^T(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*) + ||\hat{\tilde{\boldsymbol{\beta}}}_L||_1 \\ &\geq -\frac{1}{2}||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 + ||\hat{\tilde{\boldsymbol{\beta}}}_L||_1\end{aligned}$$

Now, definition of l_1 norm, fact that $\tilde{\boldsymbol{\beta}}^*_{s^{*c}} = \mathbf{0}_{|s^{*c}|}$, triangle inequality: $||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*}||_1 \geq ||\tilde{\boldsymbol{\beta}}^*||_1 - ||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*} - \tilde{\boldsymbol{\beta}}^*||_1$ and again definition of l_1 norm give:

$$\begin{aligned}||\tilde{\boldsymbol{\beta}}^*||_1 &\geq -\frac{1}{2}||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 + ||\hat{\tilde{\boldsymbol{\beta}}}_L||_1 = -\frac{1}{2}|\hat{\boldsymbol{\beta}}_{L,0} - \boldsymbol{\beta}^*_0| - \frac{1}{2}||\hat{\tilde{\boldsymbol{\beta}}}_L - \tilde{\boldsymbol{\beta}}^*||_1 + ||\hat{\tilde{\boldsymbol{\beta}}}_L||_1 \\ &= -\frac{1}{2}|\hat{\boldsymbol{\beta}}_{L,0} - \boldsymbol{\beta}^*_0| - \frac{1}{2}||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*} - \tilde{\boldsymbol{\beta}}^*_{s^*}||_1 - \frac{1}{2}||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^{*c}}||_1 + ||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*}||_1 + ||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^{*c}}||_1 \\ &= -\frac{1}{2}|\hat{\boldsymbol{\beta}}_{L,0} - \boldsymbol{\beta}^*_0| - \frac{1}{2}||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*} - \tilde{\boldsymbol{\beta}}^*_{s^*}||_1 + ||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*}||_1 + \frac{1}{2}||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^{*c}}||_1 \\ &\geq -\frac{1}{2}|\hat{\boldsymbol{\beta}}_{L,0} - \boldsymbol{\beta}^*_0| - \frac{1}{2}||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*} - \tilde{\boldsymbol{\beta}}^*_{s^*}||_1 + ||\tilde{\boldsymbol{\beta}}^*||_1 - ||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*} - \tilde{\boldsymbol{\beta}}^*||_1 + \frac{1}{2}||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^{*c}}||_1 \\ &= -\frac{1}{2}|\hat{\boldsymbol{\beta}}_{L,0} - \boldsymbol{\beta}^*_0| - \frac{3}{2}||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*} - \tilde{\boldsymbol{\beta}}^*||_1 + ||\tilde{\boldsymbol{\beta}}^*||_1 + \frac{1}{2}||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^{*c}}||_1.\end{aligned}$$

Hence, using again that $\tilde{\boldsymbol{\beta}}^*_{s^{*c}} = 0$, rearranging terms and multiplying the above inequality by 2, we obtain:

$$3||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^*} - \tilde{\boldsymbol{\beta}}^*||_1 + |\hat{\boldsymbol{\beta}}_{L,0} - \boldsymbol{\beta}^*_0| \geq ||\hat{\tilde{\boldsymbol{\beta}}}_{L,s^{*c}} - \tilde{\boldsymbol{\beta}}^*_{s^{*c}}||_1. \tag{A.21}$$

Now, we define a map $F : \mathbb{R}^{p_n+1} \to \mathbb{R}$:

$$F(\boldsymbol{\Delta}) = R_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - R_n(\boldsymbol{\beta}^*) + \lambda(||\tilde{\boldsymbol{\beta}}^* + \tilde{\boldsymbol{\Delta}}||_1 - ||\tilde{\boldsymbol{\beta}}^*||_1)$$

and sets:

$$\tilde{\mathcal{C}} = \{\boldsymbol{\Delta} \in \mathbb{R}^{p_n+1} : 3||\tilde{\boldsymbol{\Delta}}_{s^*}||_1 + |\Delta_0| \geq ||\tilde{\boldsymbol{\Delta}}_{s^{*c}}||_1\} \subseteq \mathcal{C},$$

$$\mathcal{D} = \{\boldsymbol{\Delta} \in \mathcal{C} : ||\boldsymbol{\Delta}||_2 = 5|s^*_0|^{1/2}\lambda\kappa_n^{-1}\}.$$

Let $G(u) = R_n(\boldsymbol{\beta}^* + u\boldsymbol{\Delta})$ for $u \in \mathbb{R}$. Analogously, as in the proof of Theorem 5 in Fan et al. (2014a), we obtain for $\boldsymbol{\Delta} \in \mathcal{D}$:

$$|G'''(u)| \leq K||\boldsymbol{\Delta}||_1 G''(u) \leq K \cdot 4\sqrt{|s^*_0|}||\boldsymbol{\Delta}||_2 G''(u) = zG''(u),$$

where

$$z = 4K\sqrt{|s^*_0|}||\boldsymbol{\Delta}||_2 = \frac{20K|s^*_0|\lambda}{\kappa_n}$$

as $\boldsymbol{\Delta} \in \mathcal{D}$. Moreover, $z \in [0, 1]$ from assumption on λ. Thus, from the Lemma A.52, we obtain (see also (4) in the proof of Theorem 5 in Fan et al. (2014a)):

$$G(1) - G(0) - G'(0) \geq G''(0)h(z),$$

where $h(z) = z^{-2}(e^{-z} + z - 1)$ and $h(z) \geq h(1) > 1/3$ for $z \in (0, 1]$ in view of Lemma A.53. Hence we obtain for $\boldsymbol{\Delta} \in \mathcal{D}$:

$$R_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - R_n(\boldsymbol{\beta}^*) \geq DR_n(\boldsymbol{\beta}^*)^T\boldsymbol{\Delta} + \frac{1}{3}\boldsymbol{\Delta}^T D^2 R_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta}. \tag{A.22}$$

Finally, from inequalities $||DR_n(\boldsymbol{\beta}^*)||_\infty \leq \frac{\lambda}{2}$, $||\tilde{\boldsymbol{\beta}}^* + \tilde{\boldsymbol{\Delta}}||_1 - ||\tilde{\boldsymbol{\beta}}^*||_1 \geq ||\tilde{\boldsymbol{\Delta}}_{s^{*c}}||_1 - ||\tilde{\boldsymbol{\Delta}}_{s^*}||_1$ (see (3) in the proof of Theorem 5 in Fan et al. (2014a)) and (A.22) we have for $\boldsymbol{\Delta} \in \tilde{\mathcal{C}}$:

$$\begin{aligned}
F(\boldsymbol{\Delta}) &\geq DR_n(\boldsymbol{\beta}^*)^T\boldsymbol{\Delta} + \frac{1}{3}\boldsymbol{\Delta}^T D^2R_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta} + \lambda(||\tilde{\boldsymbol{\beta}}^* + \tilde{\boldsymbol{\Delta}}||_1 - ||\tilde{\boldsymbol{\beta}}^*||_1) \\
&\geq -\frac{\lambda}{2}||\boldsymbol{\Delta}||_1 + \frac{\kappa_n}{3}||\boldsymbol{\Delta}||_2^2 + \lambda(||\tilde{\boldsymbol{\Delta}}_{s^{*c}}||_1 - ||\tilde{\boldsymbol{\Delta}}_{s^*}||_1) \\
&= \frac{\kappa_n}{3}||\boldsymbol{\Delta}||_2^2 - \frac{\lambda}{2}||\boldsymbol{\Delta}_{s_0^*}||_1 - \frac{\lambda}{2}||\tilde{\boldsymbol{\Delta}}_{s^{*c}}||_1 + \lambda||\tilde{\boldsymbol{\Delta}}_{s^{*c}}||_1 - \lambda||\tilde{\boldsymbol{\Delta}}_{s^*}||_1 \\
&= \frac{\kappa_n}{3}||\boldsymbol{\Delta}||_2^2 + \frac{\lambda}{2}||\tilde{\boldsymbol{\Delta}}_{s^{*c}}||_1 - \frac{\lambda}{2}||\boldsymbol{\Delta}_{s_0^*}||_1 - \lambda||\tilde{\boldsymbol{\Delta}}_{s^*}||_1 \\
&\geq \frac{\kappa_n}{3}||\boldsymbol{\Delta}||_2^2 - \frac{3\lambda}{2}||\boldsymbol{\Delta}_{s_0^*}||_1 \\
&\geq \frac{\kappa_n}{3}||\boldsymbol{\Delta}||_2^2 - \frac{3\lambda}{2}\sqrt{|s_0^*|}||\boldsymbol{\Delta}||_2 = \frac{5|s_0^*|\lambda^2}{6\kappa_n} > 0.
\end{aligned}$$

This ends the proof, because in a view of Lemma 4 in Negahban et al. (2012), we get $||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_2 \leq 5|s_0^*|^{1/2}\lambda\kappa_n^{-1}$ (see detailed explanation in the proof of Theorem 5 in Fan et al. (2014a)). □

A.7. Technical lemmas

Lemmas A.43 and A.44 can be used to check the sign of proportionality constant η in Remark 2.17. Lemma A.44 is a slight modification of Lemma A.43 for strictly increasing functions and has analogous proof.

Lemma A.43 (Thorisson (1995, Section 2))**.** *Let U be a random variable and $f, g\colon \mathbb{R} \to \mathbb{R}$ be non-decreasing functions. Then* $\mathrm{Cov}(f(U), g(U)) \geq 0$.

Lemma A.44. *Let U be a random variable satisfying condition $\mathbb{P}(U = c) < 1$ for all $c \in \mathbb{R}$ and $f, g\colon \mathbb{R} \to \mathbb{R}$ be strictly increasing functions. Then* $\mathrm{Cov}(f(U), g(U)) > 0$.

Proof. Since f and g are strictly increasing, then $(f(x) - f(y))(g(x) - g(y)) > 0$ for all $x, y \in \mathbb{R}$ with $x \neq y$. Let V be an independent copy of variable U. Then

$$\mathbb{P}(U = V) = \mathbb{E}I(U = V) = \mathbb{E}\mathbb{P}(U = V|V) < 1$$

from independence and we have:

$$\begin{aligned}
0 &< \mathbb{E}(f(U) - f(V))(g(U) - g(V))I(U \neq V) = \mathbb{E}(f(U) - f(V))(g(U) - g(V)) \\
&= \mathbb{E}f(U)g(U) + \mathbb{E}f(V)g(V) - \mathbb{E}f(U)g(V) - \mathbb{E}f(V)g(U) \\
&= 2\mathbb{E}f(U)g(U) - 2\mathbb{E}f(U)\mathbb{E}g(U) = 2\,\mathrm{Cov}(f(U), g(U)).
\end{aligned}$$

□

Below we state Stein's lemma which is useful in the semiparametric setup. First version of this lemma appeared in Stein (1981). Proof of the most general version for multivariate normal distribution (Lemma A.46) can be found in Liu (1994).

Lemma A.45. *(Stein's lemma for normal distribution) Suppose that the random vector* $(Z_1, Z_2)^T$ *has a bivariate normal distribution and* $f\colon \mathbb{R} \to \mathbb{R}$ *is a differentiable function fulfilling* $\mathbb{E}|f'(Z_1)| < \infty$*, then*

$$\mathrm{Cov}(f(Z_1), Z_2) = \mathrm{Cov}(Z_1, Z_2)\mathbb{E}f'(Z_1).$$

Lemma A.46. *(Stein's lemma for multivariate normal distribution) Suppose that the random vector* $(\boldsymbol{Z}_1^T, \boldsymbol{Z}_2^T)^T$*, where* $\boldsymbol{Z}_1 \in \mathbb{R}^{m_1}$, $\boldsymbol{Z}_2 \in \mathbb{R}^{m_2}$ *has a multivariate normal distribution and* $f\colon \mathbb{R}^{m_2} \to \mathbb{R}$ *is a differentiable function fulfilling* $\mathbb{E}||Df(\boldsymbol{Z}_2)||_2 < \infty$*, then*

$$\mathrm{Cov}(\boldsymbol{Z}_1, f(\boldsymbol{Z}_2)) = \mathrm{Cov}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\mathbb{E}Df(\boldsymbol{Z}_2).$$

Lemma A.47 (Hjort and Pollard (1993, Lemma 2))**.** *Suppose* $A_n\colon \mathcal{S} \to \mathbb{R}, B_n\colon \mathcal{S} \to \mathbb{R}$ *be a sequence of random functions defined on an open convex set* $\mathcal{S} \in \mathbb{R}^p$*. Assume that* A_n *are convex functions. Let* $\boldsymbol{a}_n = \arg\min_{\boldsymbol{v}\in\mathcal{S}} A_n(\boldsymbol{v})$, $\boldsymbol{b}_n = \arg\min_{\boldsymbol{v}\in\mathcal{S}} B_n(\boldsymbol{v})$ *and* $\boldsymbol{b}_n$ *is unique. Let* $\delta > 0$*. If* $||\boldsymbol{a}_n - \boldsymbol{b}_n||_2 \geq \delta$ *then*

$$\sup_{||\boldsymbol{v}-\boldsymbol{b}_n||_2\leq\delta} |A_n(\boldsymbol{v}) - B_n(\boldsymbol{v})| \geq \frac{1}{2} \inf_{||\boldsymbol{v}-\boldsymbol{b}_n||_2=\delta} B_n(\boldsymbol{v}) - B_n(\boldsymbol{b}_n).$$

Remark A.48. *Lemma A.47 is true even when* $\boldsymbol{a}_n$ *is not unique.*

Theorem A.49. *Assume that* $\boldsymbol{X}_n, \boldsymbol{X} \in \mathbb{R}^{p+1}$ *are random variables such that*

$$\mathbb{E}||\boldsymbol{X}_n - \boldsymbol{X}||_2 \to 0,$$

$\boldsymbol{X}$ *is integrable and* q *is uniformly continuous. Let* $\mathbb{P}(Y_n = 1|\boldsymbol{X}_n) = q(\boldsymbol{X}_n^T\boldsymbol{\beta})$, $\mathbb{P}(Y = 1|\boldsymbol{X}) = q(\boldsymbol{X}^T\boldsymbol{\beta})$,

$$\boldsymbol{\beta}_n^* = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{p+1}} \mathbb{E}l(\boldsymbol{b}, \boldsymbol{X}_n, Y_n)$$

and

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{p+1}} \mathbb{E}l(\boldsymbol{b}, \boldsymbol{X}, Y),$$

where l *is a logistic loss (see (1.9)). Then we have* $\boldsymbol{\beta}_n^* \to \boldsymbol{\beta}^*$*.*

Proof. Let $f_n(\mathbf{b}) = l(\mathbf{b}, \mathbf{X}_n, Y_n)$, $f(\mathbf{b}) = l(\mathbf{b}, \mathbf{X}, Y)$. We first note that the uniform convergence holds for $|\mathbb{E}f_n(\mathbf{b}) - \mathbb{E}f(\mathbf{b})|$ on bounded sets, that is, for any finite K:

$$\sup_{||\mathbf{b}||_2\leq K} |\mathbb{E}f_n(\mathbf{b}) - \mathbb{E}f(\mathbf{b})| \to 0. \tag{A.23}$$

Indeed, using definition of l, triangle inequality, the Schwarz's inequality, the mean value theorem and boundedness of q we get the following sequence of inequalities:

$$\begin{aligned}|\mathbb{E}f_n(\mathbf{b}) - \mathbb{E}f(\mathbf{b})| &\leq |\mathbb{E}(Y_n\mathbf{X}_n\mathbf{b} - Y\mathbf{X}\mathbf{b})| + \left|\mathbb{E}\left(\ln\left(1 + e^{\mathbf{X}_n^T\mathbf{b}}\right) - \ln\left(1 + e^{\mathbf{X}^T\mathbf{b}}\right)\right)\right| \\ &= |\mathbb{E}(q(\mathbf{X}_n^T\boldsymbol{\beta})\mathbf{X}_n\mathbf{b} - q(\mathbf{X}^T\boldsymbol{\beta})\mathbf{X}\mathbf{b})|\end{aligned}$$

$$
\begin{aligned}
&+ \left|\mathbb{E}\left(\ln\left(1+e^{\mathbf{X}_n^T\mathbf{b}}\right) - \ln\left(1+e^{\mathbf{X}^T\mathbf{b}}\right)\right)\right| \\
&\leq |\mathbb{E}(q(\mathbf{X}_n^T\boldsymbol{\beta})\mathbf{X}_n^T\mathbf{b} - q(\mathbf{X}^T\boldsymbol{\beta})\mathbf{X}^T\mathbf{b})| + ||\mathbf{b}||_2\mathbb{E}||\mathbf{X}_n - \mathbf{X}||_2 \\
&\leq |\mathbb{E}(q(\mathbf{X}_n^T\boldsymbol{\beta}) - q(\mathbf{X}^T\boldsymbol{\beta}))\mathbf{X}^T\mathbf{b}| + \mathbb{E}|q(\mathbf{X}_n^T\boldsymbol{\beta})||\mathbf{X}_n^T\mathbf{b} - \mathbf{X}^T\mathbf{b}| \\
&\quad + ||\mathbf{b}||_2\mathbb{E}||\mathbf{X}_n - \mathbf{X}||_2 \\
&\leq \mathbb{E}|(q(\mathbf{X}_n^T\boldsymbol{\beta}) - q(\mathbf{X}^T\boldsymbol{\beta}))\mathbf{X}^T\mathbf{b}| + 2||\mathbf{b}||_2\mathbb{E}||\mathbf{X}_n - \mathbf{X}||_2.
\end{aligned}
$$

Now, observe that from uniform continuity of q, for any $\varepsilon > 0$ exists $\delta > 0$ such that for large n if $||\mathbf{X}_n - \mathbf{X}||_2 < \delta$, then $|q(\mathbf{X}_n^T\boldsymbol{\beta}) - q(\mathbf{X}^T\boldsymbol{\beta})| < \varepsilon$ and we have for large n:

$$
\begin{aligned}
\mathbb{E}|(q(\mathbf{X}_n^T\boldsymbol{\beta}) - q(\mathbf{X}^T\boldsymbol{\beta}))\mathbf{X}^T\mathbf{b}| &\leq \varepsilon\mathbb{E}I(||\mathbf{X}_n - \mathbf{X}||_2 < \delta)|\mathbf{X}^T\mathbf{b}| + \mathbb{E}I(||\mathbf{X}_n - \mathbf{X}||_2 \geq \delta)|\mathbf{X}^T\mathbf{b}| \\
&\leq \varepsilon||\mathbf{b}||_2\mathbb{E}||\mathbf{X}||_2 + ||\mathbf{b}||_2\mathbb{E}I(||\mathbf{X}_n - \mathbf{X}||_2 \geq \delta)||\mathbf{X}||_2 \leq C\varepsilon||\mathbf{b}||_2,
\end{aligned}
$$

where C is some constant. Convergence in (A.23) readily follows from this as $\varepsilon > 0$ was arbitrary. We now prove that $\boldsymbol{\beta}_n^* \to \boldsymbol{\beta}^*$. If it does not hold then for a certain $k_n \in \mathbb{N}$, $k_n \to \infty$ we have $||\boldsymbol{\beta}_{k_n}^* - \boldsymbol{\beta}^*||_2 \geq \delta$ for some $\delta > 0$. From uniqueness of $\boldsymbol{\beta}^*$ and Lemma A.47 for $A_n = f_{k_n}$, $\mathbf{a}_n = \boldsymbol{\beta}_{k_n}^*$, $B_n = f$ and $\mathbf{b}_n = \boldsymbol{\beta}^*$ it directly follows that:

$$
\sup_{||\mathbf{b}-\boldsymbol{\beta}^*||_2 \leq \delta} |\mathbb{E}f_{k_n}(\mathbf{b}) - \mathbb{E}f(\mathbf{b})| \geq \frac{1}{2}(\mathbb{E}f(\boldsymbol{\beta}^*) - \sup_{||\mathbf{b}-\boldsymbol{\beta}^*||_2 = \delta} \mathbb{E}f(\mathbf{b})) > 0,
$$

which contradicts (A.23). □

Lemma A.50. *Let $\boldsymbol{U} \sim \mathcal{N}_k(0, I)$, $f\colon \mathbb{R} \to \mathbb{R}$, $f \in C^2$, f is bounded and $\boldsymbol{\eta} \in \mathbb{R}^k$. Then:*

$$
\mathbb{E}f(\boldsymbol{\eta}^T\boldsymbol{U})U_iU_j = \begin{cases} \mathbb{E}f''(\boldsymbol{\eta}^T\boldsymbol{U})\eta_i\eta_j & i \neq j \\ \mathbb{E}f(\boldsymbol{\eta}^T\boldsymbol{U}) + \mathbb{E}f''(\boldsymbol{\eta}^T\boldsymbol{U})\eta_i^2 & i = j \end{cases}.
$$

Proof. Let $i \neq j$. Then from Lemma A.45 we have:

$$
\begin{aligned}
\mathbb{E}f(\boldsymbol{\eta}^T\mathbf{U})U_iU_j &= \mathbb{E}(U_i\mathbb{E}(f(\boldsymbol{\eta}^T\mathbf{U})U_j|U_i)) = \mathbb{E}(U_i \cdot \mathbb{E}(f'(\boldsymbol{\eta}^T\mathbf{U})|U_i) \cdot \operatorname{Cov}(\boldsymbol{\eta}^T\mathbf{U}, U_j|U_i)) \\
&= \mathbb{E}f'(\boldsymbol{\eta}^T\mathbf{U})U_i\eta_j = \mathbb{E}f''(\boldsymbol{\eta}^T\mathbf{U})\operatorname{Cov}(\boldsymbol{\eta}^T\mathbf{U}, U_i)\eta_j = \mathbb{E}f''(\boldsymbol{\eta}^T\mathbf{U})\eta_i\eta_j.
\end{aligned}
$$

Now let $i = j$. By integration by parts and the Stein's lemma we have:

$$
\begin{aligned}
\mathbb{E}f(\boldsymbol{\eta}^T\mathbf{U})U_i^2 &= \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^{k-1}} \exp\left(-\frac{\sum_{j\neq i} u_j^2}{2}\right) \int_{\mathbb{R}} f(\boldsymbol{\eta}^T\mathbf{u})u_i \cdot u_i \exp\left(-\frac{u_i^2}{2}\right) du_i \prod_{j\neq i} du_j \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^{k-1}} \exp\left(-\frac{\sum_{j\neq i} u_j^2}{2}\right) \int_{\mathbb{R}} (f'(\boldsymbol{\eta}^T\mathbf{u})\eta_i u_i + f(\boldsymbol{\eta}^T\mathbf{u})) \exp\left(-\frac{u_i^2}{2}\right) du_i \prod_{j\neq i} du_j \\
&= \mathbb{E}f'(\boldsymbol{\eta}^T\mathbf{U})U_i\eta_i + \mathbb{E}f(\boldsymbol{\eta}^T\mathbf{U}) = \mathbb{E}f''(\boldsymbol{\eta}^T\mathbf{U})\eta_i^2 + \mathbb{E}f(\boldsymbol{\eta}^T\mathbf{U}).
\end{aligned}
$$

□

Note that statement 1 of Lemma A.51 below is similar to Lemma A.44. However, it uses different method of proof which is used also in the proof of statement 3 of Lemma A.51.

Lemma A.51. *(Expectation inequalities)*

1. *Let $a > 0$, $X \in \mathbb{R}$ be a random variable such that $\mathbb{P}(X = 0) < 1, X \in L_1, \mathbb{E}X = 0$. Let $f\colon \mathbb{R} \to \mathbb{R}$ be strictly increasing, bounded, positive function and Z - random variable independent of X. Then $\mathbb{E}f(aX + Z)X > 0$.*
2. *If $f\colon \mathbb{R} \to \mathbb{R}$ is positive, bounded function, $\boldsymbol{U} \in \mathbb{R}^p$ is a random vector, $\mathbb{E}||\boldsymbol{U}||^2 < \infty$, $\mathbb{P}(\boldsymbol{\lambda}^T \boldsymbol{U} \neq 0) > 0$ for every $\boldsymbol{\lambda} \in \mathbb{R}^p \setminus \{0\}$. Then matrix $\mathbb{E}f(\boldsymbol{\eta}^T \boldsymbol{U})\boldsymbol{U}\boldsymbol{U}^T$ is positive definite for $\boldsymbol{\eta} \in \mathbb{R}^p$.*
3. *If $f\colon \mathbb{R} \to \mathbb{R}$ is a function such that for every $x > 0$ we have $f(x) > f(0) = 0 > f(-x)$, $X \in \mathbb{R}$ is random variable such that $\mathbb{P}(X = 0) < 1$ and $\mathbb{E}|f(X)X| < \infty$, then $\mathbb{E}f(X)X > 0$.*
4. *If $X \in \mathbb{R}$ is a random variable such that $\mathbb{E}|X| < \infty$ and $\mathbb{P}(X = 0) < 1$, then $\mathbb{E}q_L''(X)X < 0$.*
5. *If $X \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, then $\mathbb{E}q_L'''(X) < 0$.*
6. *If $X \sim \mathcal{N}(0, \sigma^2)$, then the function $f(u) = u\mathbb{E}q_L'(uX)$ is strictly increasing for $u > 0$.*

Proof. 1. As f is strictly increasing and $a > 0$, we have:

$$\begin{aligned}\mathbb{E}f(aX + Z)X &= \mathbb{E}f(aX + Z)XI(X > 0) + \mathbb{E}f(aX + Z)XI(X < 0)\\ &> \mathbb{E}f(Z)XI(X > 0) + \mathbb{E}f(Z)XI(X < 0) = \mathbb{E}f(Z)X = \mathbb{E}f(Z)\mathbb{E}X = 0.\end{aligned}$$

2. For $\boldsymbol{\lambda} \in \mathbb{R}^p \setminus \{0\}$, we have:
$$\boldsymbol{\lambda}^T(\mathbb{E}f(\boldsymbol{\eta}^T\mathbf{U})\mathbf{U}\mathbf{U}^T)\boldsymbol{\lambda} = \mathbb{E}f(\boldsymbol{\eta}^T\mathbf{U})||\boldsymbol{\lambda}^T\mathbf{U}||^2 I(\boldsymbol{\lambda}^T\mathbf{U} \neq 0) > 0.$$
3. Proof is analogous to proof of 1.
4. Note that $-q_L''$ satisfies assumptions of p. 3 and is bounded. This ends the proof.
5. From Stein's lemma we have: $\mathbb{E}q_L''(X)X = \sigma^2\mathbb{E}q_L'''(X)$, hence the inequality follows from 4.
6. Observe that from the Stein's lemma we have: $f(u) = \sigma^{-2} \cdot \mathbb{E}q_L(uX)X$. Hence $f'(u) = \sigma^{-2} \cdot \mathbb{E}q_L'(uX)X^2 > 0$.

□

The following lemma (see Bach (2010)) provides inequality, which is used in Lemma A.54 and Theorem A.42 (see (A.22)) to give quadratic lower bounds respectively for risk function R and for empirical risk R_n when function ρ satisfies condition (stronger than convexity) of the form:

$$\left|\frac{\partial^3 \rho}{\partial b^3}(b, y)\right| \leq K(y)\frac{\partial^2 \rho}{\partial b^2}(b, y).$$

for all $b \in \mathbb{R}, y \in \{0, 1\}$. This condition is in particular satisfied for logistic loss (see (A.25)). Lemma A.53 is an auxiliary fact to prove inequalities (A.22) and (A.24) - we note that constant $1/3$ occurring in them in quadratic terms can be replaced by e^{-1}. Lemma A.54 shows that the assumptions regarding quadratic lower bounds of risk function in neighbourhood of $\boldsymbol{\beta}^*$ (namely (MC) and $(C_\epsilon(w))$) are reasonable.

Lemma A.52 (Bach (2010, Lemma 1))**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a three times differentiable function such that for all $t \in \mathbb{R}$: $|g'''(t)| \leq Sg''(t)$ for some $S \geq 0$. Then, for all $t \geq 0$:*

- $\frac{g''(0)}{S^2}(e^{-St} + St - 1) \leq g(t) - g(0) - g'(0)t \leq \frac{g''(0)}{S^2}(e^{St} - St - 1)$, *if* $S > 0$,
- $g(t) - g(0) - g'(0)t = \frac{g''(0)t^2}{2}$, *if* $S = 0$.

Lemma A.53.

$$\frac{1}{2} > \frac{e^{-x} + x - 1}{x^2} \geq e^{-1}$$

for $x \in (0, 1]$.

Proof. Let $f(x) = (e^{-x} + x - 1)x^{-2}$. Using inequality $e^{-x} \leq 1 - x + x^2/2$ for $x \geq 0$ (where equality holds only for $x = 0$) we obtain for $x > 0$:

$$f(x) < \frac{1 - x + x^2/2 + x - 1}{x^2} = \frac{1}{2}.$$

To prove the right inequality, we compute derivative of f:

$$f'(x) = \frac{(-e^{-x} + 1)x^2 - (e^{-x} + x - 1) \cdot 2x}{x^4} = \frac{-x + 2 - e^{-x}(x + 2)}{x^3}.$$

We now prove that $f'(x) \leq 0$ for $x \in (0, 1]$. It is enough to prove that for $x \in [0, 1]$:

$$g(x) = -e^{-x}(x + 2) - x + 2 \leq 0.$$

Using again inequality $e^{-x} \leq 1 - x + x^2/2$ for $x \geq 0$ gives:

$$g'(x) = e^{-x}(x + 2) - e^{-x} - 1 = e^{-x}(x + 1) - 1 \leq \left(1 - x + \frac{x^2}{2}\right)(x + 1) - 1 = \frac{1}{2}(x - 1)x^2.$$

Hence $g'(x) \leq 0$ for $x \in [0, 1]$. This means that $g(x) \leq g(0) = 0$ and $f'(x) = g(x)x^{-3} \leq 0$ for $x \in (0, 1]$. Thus function f is decreasing and we obtain:

$$\frac{e^{-x} + x - 1}{x^2} = f(x) \geq f(1) = e^{-1}.$$

This ends the proof. □

Lemma A.54. *If $R : \mathbb{R}^{p_n+1} \to \mathbb{R}$ is a risk function (see (1.2)) for logistic loss defined in (1.9), $\boldsymbol{X} \in \mathbb{R}^{p_n+1}$ is bounded random vector: $||\boldsymbol{X}||_\infty \leq M$ $\mathbb{P}_{\boldsymbol{X}}$ a.e., then for any $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^{p_n+1}$ with $||\boldsymbol{b}_1 - \boldsymbol{b}_2||_1 \leq 1/M$ we have:*

$$R(\boldsymbol{b}_1) \geq R(\boldsymbol{b}_2) + DR(\boldsymbol{b}_2)^T(\boldsymbol{b}_1 - \boldsymbol{b}_2) + \frac{1}{3}(\boldsymbol{b}_1 - \boldsymbol{b}_2)^T D^2 R(\boldsymbol{b}_2)(\boldsymbol{b}_1 - \boldsymbol{b}_2). \tag{A.24}$$

Proof. Firstly we observe that if $\rho(b, y) = -yb + \ln\left(1 + e^b\right)$, then we have:

$$\left|\frac{\partial^3 \rho}{\partial b^3}(b, y)\right| = \frac{e^b}{(1 + e^b)^2}\left|\frac{1 - e^b}{1 + e^b}\right| \leq \frac{e^b}{(1 + e^b)^2} = \frac{\partial^2 \rho}{\partial b^2}(b, y). \tag{A.25}$$

Let $t \in \mathbb{R}$ and $\tilde{R}(t) = R(\mathbf{b}_2 + t(\mathbf{b}_1 - \mathbf{b}_2))$. We calculate that:

$$\tilde{R}''(t) = \mathbb{E}\frac{\partial^2 \rho}{\partial b^2}(\mathbf{b}_2^T\mathbf{X} + t(\mathbf{b}_1 - \mathbf{b}_2)^T\mathbf{X}, Y)((\mathbf{b}_1 - \mathbf{b}_2)^T\mathbf{X})^2,$$

$$\tilde{R}'''(t) = \mathbb{E}\frac{\partial^3 \rho}{\partial b^3}(\mathbf{b}_2^T\mathbf{X} + t(\mathbf{b}_1 - \mathbf{b}_2)^T\mathbf{X}, Y)((\mathbf{b}_1 - \mathbf{b}_2)^T\mathbf{X})^3.$$

Hence in view of (A.25) and inequality $|(\mathbf{b}_1-\mathbf{b}_2)^T\mathbf{X}| \leq ||\mathbf{X}||_\infty ||\mathbf{b}_1-\mathbf{b}_2||_1 \leq M||\mathbf{b}_1-\mathbf{b}_2||_1$ we obtain:

$$\begin{aligned}|\tilde{R}'''(t)| &\leq \mathbb{E}\left|\frac{\partial^3\rho}{\partial b^3}(\mathbf{b}_2^T\mathbf{X}+t(\mathbf{b}_1-\mathbf{b}_2)^T\mathbf{X},Y)\right|\left|(\mathbf{b}_1-\mathbf{b}_2)^T\mathbf{X}\right|^3\\ &\leq M||\mathbf{b}_1-\mathbf{b}_2||_1\mathbb{E}\frac{\partial^2\rho}{\partial b^2}(\mathbf{b}_2^T\mathbf{X}+t(\mathbf{b}_1-\mathbf{b}_2)^T\mathbf{X},Y)\left|(\mathbf{b}_1-\mathbf{b}_2)^T\mathbf{X}\right|^2\\ &= M||\mathbf{b}_1-\mathbf{b}_2||_1\tilde{R}''(t). \end{aligned} \tag{A.26}$$

Using Lemma A.52 for $g=\tilde{R}$ and $S=M||\mathbf{b}_1-\mathbf{b}_2||_1$ yields for $t\geq 0$:

$$\tilde{R}(t) \geq \tilde{R}(0)+\tilde{R}'(0)t+\tilde{R}''(0)\cdot\left(\frac{e^{-M||\mathbf{b}_1-\mathbf{b}_2||_1 t}+M||\mathbf{b}_1-\mathbf{b}_2||_1 t-1}{M^2||\mathbf{b}_1-\mathbf{b}_2||_1^2}\right).$$

Using definition of $\tilde{R}$, above inequality can be rewritten for $t=1$ as:

$$\begin{aligned}R(\mathbf{b}_1) \geq\ & R(\mathbf{b}_2)+DR(\mathbf{b}_2)^T(\mathbf{b}_1-\mathbf{b}_2)\\ &+(\mathbf{b}_1-\mathbf{b}_2)^T D^2R(\mathbf{b}_2)(\mathbf{b}_1-\mathbf{b}_2)\cdot\left(\frac{e^{-M||\mathbf{b}_1-\mathbf{b}_2||_1}+M||\mathbf{b}_1-\mathbf{b}_2||_1-1}{M^2||\mathbf{b}_1-\mathbf{b}_2||_1^2}\right).\end{aligned}$$

Now, in view of inequality $M||\mathbf{b}_1-\mathbf{b}_2||_1\leq 1$, which follows from assumptions and Lemma A.53 we have:

$$\frac{e^{-M||\mathbf{b}_1-\mathbf{b}_2||_1}+M||\mathbf{b}_1-\mathbf{b}_2||_1-1}{M^2||\mathbf{b}_1-\mathbf{b}_2||_1^2} \geq e^{-1} > \frac{1}{3}.$$

This ends the proof, as

$$(\mathbf{b}_1-\mathbf{b}_2)^T D^2R(\mathbf{b}_2)(\mathbf{b}_1-\mathbf{b}_2) = \mathbb{E}q_L(\mathbf{b}_2^T\mathbf{X})(1-q_L(\mathbf{b}_2^T\mathbf{X}))((\mathbf{b}_1-\mathbf{b}_2)^T\mathbf{X})^2 > 0.$$

□

Lemma A.55. *If $R:\mathbb{R}^{p_n+1}\to\mathbb{R}$ is a risk function (see (1.2)) for quadratic loss defined in (1.11), $\boldsymbol{X}\in\mathbb{R}^{p_n+1}$ is a random vector such that $\mathbb{E}||\boldsymbol{X}||_2^2\leq\infty$, then for any $\boldsymbol{b}_1,\boldsymbol{b}_2\in\mathbb{R}^{p_n+1}$ we have:*

$$R(\boldsymbol{b}_1)-R(\boldsymbol{b}_2) = DR(\boldsymbol{b}_2)^T(\boldsymbol{b}_1-\boldsymbol{b}_2)+\frac{1}{2}(\boldsymbol{b}_1-\boldsymbol{b}_2)^T D^2R(\boldsymbol{b}_2)(\boldsymbol{b}_1-\boldsymbol{b}_2). \tag{A.27}$$

Proof. Proof follows from Taylor's expansion after noting that

$$R(\mathbf{b}) = \frac{1}{2}\mathbf{b}^T\mathbb{E}\mathbf{X}\mathbf{X}^T\mathbf{b}-\mathbf{b}^T\mathbf{X}Y+Y^2$$

and thus derivative of the order higher than 2 disappears. □

Bibliography

Alzer, H. (2010). Error function inequalities. *Advances in Computational Mathematics* **33**(3): 349–379.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics* **4**: 384–414.

Baraniuk, R., Davenport, M. A., Duarte, M. F., Hegde, C., Laska, J., Sheikh, M. and Yin, W. (2011). An Introduction to Compressive Sensing. Available from: `http://engold.ui.ac.ir/~sabahi/An%20Introduction%20to%20Compressive%20Sensing.pdf`.

Bateman, H. (1953). *Higher Transcendental Functions [Volumes I-III]*. McGraw-Hill Book Company.

Beck, A. (2014). *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*, vol. 19. SIAM.

Birnbaum, Z. W. (1942). An inequality for Mill's ratio. *The Annals of Mathematical Statistics* **13**(2): 245–246.

Brillinger, D. R. (1982). A generalized linear model with 'gaussian' regressor variables. In: *A Festschrift For Erich L. Lehmann*, CRC Press, pp. 97–114.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series Theory and Methods*. Springer.

Bühlmann, P., van de Geer, S. et al. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics* **9**(1): 1449–1473.

Cambanis, S., Huang, S. and Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis* **11**(3): 368–385.

Eaton, M. L. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis* **20**(2): 272–276.

Fan, J., Xue, L. and Zou, H. (2014a). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics* **42**(3): 819–849.

Fan, J., Xue, L. and Zou, H. (2014b). Supplement to "Strong oracle optimality of folded concave penalized estimation.".

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(3): 531–552.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1): 1–22.

Gail, M., Tan, W.-Y. and Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika* **75**(1): 57–64.

Gordon, R. D. (1941). Values of Mill's ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics* **12**(3): 364–366.

Hardin, C. D. (1982). On the linearity of regression. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **61**(3): 293–302.

Hjort, N. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. *Unpublished manuscript* Available from: `http://www.stat.yale.edu/pollard/Papers/convex.pdf`.

Huang, J., Ma, S. and Zhang, C.-H. (2008). The iterated lasso for high-dimensional logistic regression. *preprint* .

Kim, Y. and Jeon, J.-J. (2016). Consistent model selection criteria for quadratically supported risks. *The Annals of Statistics* **44**(6): 2467–2496.

Kubkowski, M. and Mielniczuk, J. (2017). Active sets of predictors for misspecified logistic regression. *Statistics* **51**(5): 1023–1045.

Kubkowski, M. and Mielniczuk, J. (2018). Projections of a general binary model on a logistic regression. *Linear Algebra and its Applications* **536**: 152–173.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics* **17**(3): 1009–1052.

Liu, J. S. (1994). Siegel's formula via Stein's identities. *Statistics & Probability Letters* **21**(3): 247–251.

Lu, W., Goldberg, Y. and Fine, J. (2012). On the robustness of the adaptive lasso to model misspecification. *Biometrika* **99**(3): 717–731.

Massart, P. (2007). *Concentration Inequalities and Model Selection*. Springer.

Mielniczuk, J. and Teisseyre, P. (2016). What do we choose when we err? Model selection and testing for misspecified logistic regression revisited. In: *Challenges in Computational Statistics and Data Mining*, Springer, pp. 271–296.

Minc, H. and Sathre, L. (1964). Some inequalities involving $(r!)^{1/r}$. *Proceedings of the Edinburgh Mathematical Society* **14**(1): 41–46.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27**(4): 538–557.

Pokarowski, P. and Mielniczuk, J. (2015). Combined ℓ_1 and greedy ℓ_0 penalized least squares for linear model selection. *Journal of Machine Learning Research* **16**(5): 961–992.

Pokarowski, P., Prochenka, A., Frej, M., Rejchel, W. and Mielniczuk, J. (2018). Improving Lasso for Model Selection and Prediction. *Unpublished manuscript* Available from: `https://www.univie.ac.at/seam/inference2018/abstracts/contributed/rejchel.pdf`.

Robbins, H. (1955). A remark on Stirling's formula. *The American Mathematical Monthly* **62**(1): 26–29.

Roberts, A. and Varberg, D. (1973). *Convex Functions*. Academic Press, New York.

Rockafellar, R. T. (1970). *Convex Analysis*, vol. 28. Princeton University Press.

Rosset, S., Zhu, J. and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research* **5**(Aug): 941–973.

Ruud, P. A. (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica* **51**(1): 225–228.

Schneider, U. and Ewald, K. (2017). On the distribution and model selection properties of the lasso estimator in low and high dimensions. *arXiv:1708.09608v1* .

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**(6): 1135–1151.

Thorisson, H. (1995). Coupling methods in probability theory. *Scandinavian Journal of Statistics* **22**(2): 159–182.

Tibshirani, R. and Wasserman, L. (2015). Sparsity and the lasso. Available from: `https://www.stat.cmu.edu/~larry/=sml/sparsity.pdf`.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics* **7**: 1456–1490.

van de Geer, S. (2016). *Estimation and Testing Under Sparsity*, vol. 2159. Springer.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y. C. and Kutyniok, G. (Editors), *Compressed Sensing: Theory and Applications*, Cambridge University Press, p. 210–268.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**(2): 307–333.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**(1): 1–25.

Yi, C. and Huang, J. (2017). Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics* **26**(3): 547–557. URL `https://doi.org/10.1080/10618600.2016.1256816`.

Zhou, S. (2010). Thresholded Lasso for high dimensional variable selection and statistical estimation. *arXiv:1002.1583* .