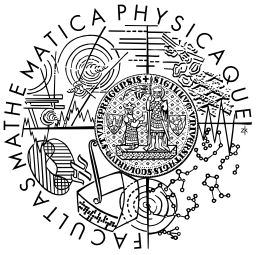


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Samir Bessisso

Řídké regresní modely

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Matúš Maciak, Ph.D.

Studijní program: Pravděpodobnost, matematická statistika a ekonometrie

Studijní obor: MPSP

Praha 2024

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Děkuji své rodině, docentu Maciakovi za vedení práce, katedře pravděpodobnosti a matematické statistiky a Matematicko-fyzikální fakultě.

Název práce: Řídké regresní modely

Autor: Bc. Samir Bessisso

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Matúš Maciak, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V řídkých lineárních regresních modelech je efekt majority vysvětlujících proměnných na podmíněnou střední hodnotu odezvy nulový. Odhady vyprodukované metodou adaptivní lasso jsou řídké a mají věštecké vlastnosti, čili asymptoticky přesně identifikují množinu nulových složek vektoru regresních koeficientů a jsou $\sqrt{n}$-konzistentními odhady nenulových regresních koeficientů. V první kapitole této diplomové práce jsou zopakovány vlastnosti odhadu metodou obyčejných nejmenších čtverců a uvedeny argumenty pro využití vychýlených regularizovaných odhadů. V druhé a třetí kapitole se zabýváme metodami lasso a adaptivní lasso. Ve čtvrté, závěrečné kapitole je diskutována problematika statistické inference po výběru rysů a odvozena metoda ke konstrukci přesných intervalových odhadů v lineárním regresním modelu, jehož množina vysvětlujících proměnných byla zvolena jako množina aktivních složek odhadu metodou lasso.

Klíčová slova: Regresní model, Regularizace, Odhadování s penalizací, Řídké odhady

Title: Sparse regression model

Author: Bc. Samir Bessisso

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Matúš Maciak, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In sparse linear regression models, the effect of the majority of explanatory variables on the conditional expected value of the response is null. The estimates produced by the adaptive lasso method are sparse and possess the oracle properties; meaning they provide asymptotically accurate identification of null elements within the regression coefficients vector while also being $\sqrt{n}$-consistent estimates of the non-zero regression coefficients. In the first chapter of this diploma thesis, we revise the properties of the ordinary least squares estimate and we present arguments favoring the adoption of biased regularized estimates. In the second and third chapters, we examine the lasso and adaptive lasso methods. In the fourth and concluding chapter of this diploma thesis, we discuss the challenges of the post-model-selection inference and we derive a method for constructing exact confidence intervals in a linear regression model whose set of the explanatory variables was chosen as a support of the lasso estimate.

Keywords: Regression model, Regularization, Penalized estimation, Sparse estimates

Obsah

Konvence a značení **2**

Úvod **3**

1 Lineární regresní metody **4**
1.1 Lineární regresní model 4
1.2 Metoda obyčejných nejmenších čtverců 6
1.3 Výběr rysů 8
1.4 Regularizace 13

2 Metoda lasso **18**
2.1 Existence, tvar a jednoznačnost 18
2.2 Výpočetní algoritmus 21
2.3 Geometrický a Bayesovský pohled 23
2.4 Asymptotické vlastnosti 30

3 Adaptivní lasso **33**
3.1 Existence, tvar a jednoznačnost 33
3.2 Výpočetní algoritmus 34
3.3 Geometrický a Bayesovský pohled 34
3.4 Asymptotické vlastnosti 36

4 Statistická inference pro lasso **37**
4.1 Statistická inference po výběru rysů 37
4.2 Přesná inference po výběru rysů 39
4.3 Kovarianční test 41

Závěr **43**

A Appendix **44**
A.1 Tvrzení o konvexních funkcích 44
A.2 Algoritmus LARS for the lasso path 45

Seznam použité literatury **46**

Seznam obrázků **49**

Konvence a značení

Konvence

- Vektory jsou sloupcové.
- Náhodné vektory jsou reálné.
- Výroky o náhodných vektorech platí ve smyslu skoro jistě.

Značení

- Pro $p \in \mathbb{N}$, $\gamma > 0$ a vektor $\boldsymbol{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ definujeme

$$\|\boldsymbol{x}\|_\gamma = \left(\sum_{j=1}^{p} |x_j|^\gamma \right)^{1/\gamma}.$$

  Pro $\gamma \geq 1$ budeme $\|\boldsymbol{x}\|_\gamma$ nazývat ℓ^γ *normou vektoru* $\boldsymbol{x}$.

- Pro $p \in \mathbb{N}$ a vektor $\boldsymbol{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ definujeme

$$\|\boldsymbol{x}\|_0 = \sum_{j=1}^{p} \mathbb{I}[x_j \neq 0],$$

  kde $\mathbb{I}[\bullet]$ je charakteristická funkce.

- Nechť $\mathcal{M}$ je množina a $\mathcal{A} \subseteq \mathcal{M}$. *Doplněk množiny* $\mathcal{A}$ *vzhledem k* $\mathcal{M}$ značíme $\mathcal{A}^C = \mathcal{M} \setminus \mathcal{A}$, případně $-\mathcal{A} = \mathcal{A}^C$.
- Nechť $\mathcal{A}$ je množina. Počet prvků množiny $\mathcal{A}$ budeme značit $|\mathcal{A}|$.
- Nechť $\mathbb{A} \in \mathbb{R}^{n \times n}$ je čtvercová matice. Stopu matice $\mathbb{A}$ budeme značit $\mathrm{tr}(\mathbb{A})$.
- **Značení pro konvergenci náhodných vektorů:** Nechť $\{\boldsymbol{X}_n\}_{n \in \mathbb{N}}$ je posloupnost reálných náhodných vektorů. Zavádíme následující značení pro konvergenci posloupnosti náhodných vektorů:
  1. Konvergence v distribuci: $\boldsymbol{X}_n \xrightarrow{\mathcal{D}} \boldsymbol{X}$, $n \to \infty$.
  2. Konvergence v pravděpodobnosti: $\boldsymbol{X}_n \xrightarrow{\mathbb{P}} \boldsymbol{X}$, $n \to \infty$.
  3. Konvergence skoro jistě: $\boldsymbol{X}_n \xrightarrow{s.j.} \boldsymbol{X}$, $n \to \infty$.
- Nechť $\widehat{\boldsymbol{\theta}}_n = T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \in \mathbb{R}^p$ je odhad parametru $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, kde $T : \mathbb{R}^n \to \mathbb{R}^p$ je měřitelná funkce. O odhadu $\widehat{\boldsymbol{\theta}}_n$ řekneme, že je $\sqrt{n}$*-konzistentním odhadem parametru* $\boldsymbol{\theta}$, jestliže

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} = O_P(1/\sqrt{n}),$$

  tedy

$$\forall \epsilon > 0 \, \exists K < \infty \, \exists n_0 \in \mathbb{N} : \sup_{n \geq n_0} \mathbb{P}(\sqrt{n} \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|_2 > K) < \epsilon.$$

Úvod

V řídkých lineárních regresních modelech je efekt majority vysvětlujících proměnných na střední hodnotu odezvy nulový nebo zanedbatelný. Je-li v řídkém lineárním regresním modelu cílem statistické analýzy vyhodnocení efektu vysvětlujících proměnných na střední hodnotu odezvy, je vhodné provést nejen odhad vektoru regresních koeficientů, ale i identifikaci signifikantních a insignifikantních složek. Běžně používaná metoda obyčejných nejmenších čtverců produkuje efektivní odhady vektoru regresních koeficientů. Využití metody obyčejných nejmenších čtverců k odhadu vektoru regresních koeficientů ovšem není vhodné v případě, kdy počet vysvětlujících proměnných převyšuje počet pozorování. Využití metody obyčejných nejmenších čtverců k identifikaci signifikantních složek vektoru regresních koeficientů vyžaduje dodatečné intervence ze strany statistika, například ve formě testování signifikance vybrané podmnožiny složek. V této diplomové práci se zabýváme metodou adaptivní lasso. Metoda adaptivní lasso produkuje řídké odhady a má věštecké vlastnosti, čili asymptoticky přesně identifikuje množinu nulových složek vektoru regresních koeficientů a je $\sqrt{n}$-konzistentním odhadem nenulových složek vektoru regresních koeficientů.

V první kapitole této diplomové práce jsou zopakovány vlastnosti odhadu metodou obyčejných nejmenších čtverců v lineárním regresním modelu a motivováno využití vychýlených regularizovaných odhadů. Dále uvádíme argumenty, proč v řídkém lineárním regresním modelu volit z ℓ^γ-regularizovaných metod, $\gamma \geq 0$, právě ℓ^1-regularizaci, čili metodu lasso. V druhé kapitole se zabýváme odhadem metodou lasso, pro který odvodíme existenci, tvar a podmínky pro jednoznačnost a uvedeme algoritmus pro výpočet. Nahlédneme na metodu lasso z geometrické perspektivy a z perspektivy Bayesovské statistiky. V závěru druhé kapitoly se věnujeme asymptotickým vlastnostem odhadu metodou lasso. Metoda lasso s kladnou pravděpodobností asymptoticky neidentifikuje množinu nulových složek vektoru regresních koeficientů, což je motivací k zavedení odhadu metodou adaptivní lasso. Třetí kapitola je věnována metodě adaptivní lasso, která již množinu nulových složek vektoru regresních koeficientů asymptoticky identifikuje. Odhad metodou adaptivní lasso lze získat pomocí metody lasso, aplikujeme-li vhodnou transformaci na matici modelu a vhodnou zpětnou transformaci na výsledný odhad metodou lasso. Díky tomuto se většina vlastností metody lasso přenáší na metodu adaptivní lasso. Ve čtvrté kapitole této diplomové práce je diskutována problematika statistické inference po výběru rysů[1], čili výběru podmnožiny vysvětlujících proměnných. Metoda (adaptivní) lasso produkuje řídké odhady a lze jí využít k automatickému výběru rysů v lineárním regresním modelu. V takovém případě je ale množina vysvětlujících proměnných zvolena na základě pozorovaných dat a klasické metody statistické inference již nejsou zcela validní. Pro metodu lasso uvedeme přesné rozdělení odhadu metodou obyčejných nejmenších čtverců v normálním lineárním regresním modelu, jehož vysvětlující proměnné byly zvoleny na základě aktivních složek odhadu metodou lasso.

Jako hlavní příspěvek této diplomové práce vnímáme shrnutí a zpracování určité problematiky na základě několika různých pramenů.

[1] Anglicky: *feature selection*

1. Lineární regresní metody

V této kapitole je v sekcích 1.1, 1.2 a 1.3 zopakován lineární regresní model, metoda obyčejných nejmenších čtverců a základní metody výběru rysů, pomocí kterých lze docílit redukce střední čtvercové chyby. V sekcích 1.1 až 1.3 čerpáme zejména z Kulich (2022), Komárek (2021), Lachout (2020), Barto a Tůma (2019) a Hastie a kol. (2009). Snaha o automatický výběr rysů a redukci střední čtvercové chyby nás dovede k regularizovaným odhadům, kterým je věnována sekce 1.4.

1.1 Lineární regresní model

Nechť Y je generická náhodná veličina a $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ je p-rozměrný generický náhodný vektor, $p \in \mathbb{N}$. Uvažujme situaci, kdy máme za úkol předpovědět hodnotu Y na základě hodnot vektoru $\boldsymbol{X}$ nebo vyhodnotit efekt náhodné veličiny X_j, $j \in \{1, \ldots, p\}$, na hodnotu Y. Náhodná veličina Y je v tomto kontextu nazývána *odezva* a náhodné veličiny $X_1, \ldots, X_p$ jsou nazývány *vysvětlující proměnné*. Jedním z přístupů, jak výše popsaný úkol řešit, je pozorovat realizace náhodného vektoru $(Y, \boldsymbol{X}^\top)^\top$ a o budoucí asociaci mezi odezvou a vysvětlujícími proměnnými činit logické indukce na základě pozorování předchozích. Nechť

$$(Y_i, X_{i,1}, \ldots, X_{i,p})^\top, \qquad i = 1, \ldots, n, \qquad n \in \mathbb{N}, \tag{1.1}$$

je náhodný výběr z pravděpodobnostního rozdělení generického náhodného vektoru $(Y, \boldsymbol{X}^\top)^\top$. Úsudky o asociaci mezi odezvou a vysvětlujícími proměnnými budeme činit pomocí statistických metod, které umožňují predikovat hodnotu odezvy na základě hodnot vysvětlujících proměnných, konstruovat efektivní odhady vlivu vysvětlujících proměnných na hodnotu odezvy a formálně kvantifikovat míru nejistoty těchto predikcí a odhadů. Použití mnohých statistických metod ale má svou cenu. O rozdělení náhodného vektoru $(Y, \boldsymbol{X}^\top)^\top$ je třeba předpokládat, že pochází z nějaké třídy pravděpodobnostních rozdělení, která se nazývá *statistický model*. Statistický model aproximuje a matematicky idealizuje mechanismus, který generuje pozorovaná data. Jaký statistický model ale volit?

Klasickým výsledkem teorie pravděpodobnosti je, že podmíněná střední hodnota $\mathbb{E}(Y \mid \boldsymbol{X})$ je ortogonální projekcí náhodné veličiny $Y \in L_2(\boldsymbol{\Omega}, \mathcal{F}, \mathbb{P})$ na prostor $L_2(\boldsymbol{X}) = L_2(\boldsymbol{\Omega}, \sigma(\boldsymbol{X}), \mathbb{P}|_{\sigma(\boldsymbol{X})})$, indukovaný náhodným vektorem $\boldsymbol{X}$

$$\mathbb{E}(Y \mid \boldsymbol{X}) = \operatorname*{argmin}_{\widetilde{\boldsymbol{X}} \in L_2(\boldsymbol{X})} \mathbb{E}(\boldsymbol{Y} - \widetilde{\boldsymbol{X}})^2$$

a podmíněná střední hodnota $\mathbb{E}(Y \mid \boldsymbol{X})$ je nejlepší aproximací náhodné veličiny Y vzhledem k střední čtvercové chybě, při známých hodnotách náhodného vektoru $\boldsymbol{X}$. Funkce $f(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^p$, je proto vhodným kandidátem na odhad odezvy Y při známých hodnotách vysvětlujících proměnných $\boldsymbol{X}$.

Funkce $f(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ se nazývá *regresní funkce*. Regresní funkce $\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ je neznámá a k řešení našeho úkolu je třeba ji odhadnout. V této práci budeme o náhodném vektoru $(Y, \boldsymbol{X}^\top)^\top$ předpokládat, že splňuje *lineární regresní model*. Lineární regresní model je parametrický model s afinní regresní funkcí, jehož parametry určují první obecný a druhý centrální moment podmíněného rozdělení $Y \mid \boldsymbol{X}$. Jedná se o základní regresní model, který umožňuje snadný a účinný odhad (lineární aproximace skutečné) regresní funkce.

Definice 1.1 (Lineární regresní model)**.** *O náhodném vektoru* $(Y, \boldsymbol{X}^\top)^\top$ *řekneme, že splňuje lineární regresní model, jestliže*

$$\mathbb{E}(Y \mid \boldsymbol{X}) = \beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta} \qquad a \qquad \mathrm{var}(Y \mid \boldsymbol{X}) = \sigma^2, \tag{1.2}$$

kde $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ *a* $\sigma^2 \in (0, \infty)$ *jsou neznámé parametry.*

Vektor $(\beta_0, \boldsymbol{\beta}^\top)^\top$ se nazývá *vektor regresních koeficientů*, σ^2 se nazývá *reziduální rozptyl*. Vlastnost $\mathrm{var}(Y \mid \boldsymbol{X}) = \sigma^2$ se nazývá *homoskedasticita*. V případě $Y \mid \boldsymbol{X} \sim \mathcal{N}(\beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta}, \sigma^2)$ budeme hovořit o *normálním lineárním regresním modelu*. Nechť $\epsilon = Y - \mathbb{E}(Y \mid \boldsymbol{X})$. Náhodná veličina ϵ se nazývá *náhodná chyba*. Platí $\mathbb{E}(\epsilon) = \mathbb{E}(\mathbb{E}[\epsilon \mid \boldsymbol{X}]) = 0$ a $\mathrm{var}(\epsilon) = \mathbb{E}(\mathrm{var}[\epsilon \mid \boldsymbol{X}]) + \mathrm{var}(\mathbb{E}[\epsilon \mid \boldsymbol{X}]) = \sigma^2$. Pomocí náhodných chyb lze model (1.2) ekvivalentně formulovat jako $Y = \beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta} + \epsilon$.

K předpokladům lineárního regresního modelu budeme navíc předpokládat, že existují smíšené momenty druhého řádu vysvětlujících proměnných

$$\mathbb{E}\left(|X_j X_k|\right) \leq \infty, \qquad j, k = 1, \ldots, p, \tag{1.3}$$

že matice $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)$ je pozitivně definitní a že sdružené pravděpodobnostní rozdělení odezvy Y a vysvětlujících proměnných $X_1, \ldots, X_p$ je spojité.

Označme $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ a $\boldsymbol{X}_j = (X_{1,j}, \ldots, X_{n,j})^\top$, $j = 1, \ldots, p$, vektory pozorovaných hodnot, $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p) \in \mathbb{R}^{n \times p}$ *matici modelu*, $\mathbb{I}_n \in \mathbb{R}^{n \times n}$ jednotkovou matici a $\mathbf{1}_n = (1, 1, \ldots, 1)^\top \in \mathbb{R}^n$. Bez újmy na obecnosti budeme předpokládat, že pozorované vysvětlující proměnné byly studentizovány, tedy

$$\boldsymbol{X}_j = \sqrt{n-1} \left\| \left(\mathbb{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top\right) \boldsymbol{X}_j \right\|_2^{-1} \cdot \left(\mathbb{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top\right) \boldsymbol{X}_j, \; j = 1, \ldots, p. \tag{1.4}$$

Metody, kterými se v této práci budeme zabývat, odhadují vektoru regresních koeficientů pomocí ortogonální projekce pozorovaných hodnot odezvy $\boldsymbol{Y}$ na zvolenou podmnožinu lineárního prostoru $\mathrm{Im}(\mathbf{1}_n, \mathbb{X}) = \{(\mathbf{1}_n, \mathbb{X})\boldsymbol{\gamma} : \boldsymbol{\gamma} \in \mathbb{R}^{p+1}\}$. Z (1.4) plyne $\mathrm{Im}(\mathbb{X}) \subseteq \mathrm{Im}(\mathbb{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)$. Zřejmě $\boldsymbol{Y} = (\mathbb{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\boldsymbol{Y} + n^{-1}\mathbf{1}_n\mathbf{1}_n^\top\boldsymbol{Y}$, kde $(\mathbb{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\boldsymbol{Y}$ je ortogonální projekcí vektoru $\boldsymbol{Y}$ na lineární prostor $\mathrm{Im}(\mathbb{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)$ a vektor $n^{-1}\mathbf{1}_n\mathbf{1}_n^\top\boldsymbol{Y}$ je ortogonální projekcí $\boldsymbol{Y}$ na lineární prostor $\mathrm{LO}\{\mathbf{1}_n\} = \{c\mathbf{1}_n : c \in \mathbb{R}\}$. Navíc jsou prostory $\mathrm{Im}(\mathbb{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)$ a $\mathrm{LO}\{\mathbf{1}_n\}$ na sebe kolmé. Budeme-li tedy pracovat s centrovaným vektorem odezev $(\mathbb{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\boldsymbol{Y}$, regresní koeficient β_0 bude projekčními metodami odhadnut jako nulový a regresní koeficienty $\beta_1, \ldots, \beta_p$ budou stále odhadnuty pomocí ortogonální projekce vektoru odezev $\boldsymbol{Y}$ na zvolenou podmnožinu prostoru $\mathrm{Im}(\mathbb{X})$. Nadále proto budeme bez újmy na obecnosti předpokládat

$$\boldsymbol{Y} = (\mathbb{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)\boldsymbol{Y} \tag{1.5}$$

a uvažovat $\beta_0 = 0$. Transformovaná data (1.4) a (1.5) splňují

$$\mathbb{E}[\boldsymbol{Y} \mid \mathbb{X}] = \mathbb{X}\boldsymbol{\beta}^* \qquad \text{a} \qquad \mathrm{Var}(\boldsymbol{Y} \mid \mathbb{X}) = \sigma^2\mathbb{I}_n. \tag{1.6}$$

Vektorem regresních koeficientů, resp. reziduálním rozptylem budeme nadále mínit neznámý vektor $\boldsymbol{\beta}^* \in \mathbb{R}^p$ resp. neznámé $\sigma^2 \in (0, \infty)$ z (1.6).

Z (1.2) je patrné, že vektor regresních koeficientů $\boldsymbol{\beta}^*$ v lineárním regresním modelu určuje asociaci mezi vysvětlujícími proměnnými a podmíněnou střední hodnotou odezvy. Vektor regresních koeficientů je ale neznámý a tak je třeba jej odhadnout. V následující sekci 1.2 je zaveden odhad vektoru regresních koeficientů metodou obyčejných nejmenších čtverců.

1.2 Metoda obyčejných nejmenších čtverců

Cílem lineární regrese je odhad vektoru regresních koeficientů, který lze využít k vyhodnocení efektu jednotlivých vysvětlujících proměnných na podmíněnou střední hodnotu odezvy nebo k predikci odezvy na základě hodnot vysvětlujících proměnných. Volba metody ke konstrukci odhadu závisí na stanoveném cíli a na struktuře dat. V lineárním regresním modelu je nejpopulárnějším odhadem vektoru regresních koeficientů *odhad metodou obyčejných nejmenších čtverců.*

Definice 1.2 (Odhad metodou obyčejných nejmenších čtverců)**.** *Uvažujme lineární regresní model (1.2). Nechť* n, $p \in \mathbb{N}$, $\boldsymbol{Y} \in \mathbb{R}^n$ *je vektor odezev a* $\mathbb{X} \in \mathbb{R}^{n \times p}$ *je matice modelu. Libovolné řešení úlohy*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2, \tag{1.7}$$

budeme nazývat odhad vektoru regresních koeficientů metodou obyčejných nejmenších čtverců a značit $\widehat{\boldsymbol{\beta}}^{OLS}$.

Dle Gaussovy–Markovovy věty (Komárek (2021), Theorem 2.4) dosahuje odhad vektoru regresních koeficientů metodou obyčejných nejmenších čtverců nejnižší střední čtvercové chyby, podmíněné pozorovanými vysvětlujícími proměnnými $\mathbb{X}$, mezi všemi nestrannými lineárními odhady vektoru $\boldsymbol{\beta}^*$.

Poznámka. Pro libovolné řešení $\widehat{\boldsymbol{\beta}}^{OLS}$ je vektor $\mathbb{X}\widehat{\boldsymbol{\beta}}^{OLS}$ jednoznačnou ortogonální projekcí odezvy $\boldsymbol{Y}$ na lineární prostor $\mathrm{Im}(\mathbb{X}) = \{\mathbb{X}\boldsymbol{\gamma} \in \mathbb{R}^n : \boldsymbol{\gamma} \in \mathbb{R}^p\}$ generovaný sloupci matice modelu (důsledek Barto a Tůma (2019), Věta 8.60 a Věta 8.61).

Důvodem k popularitě odhadu metodou obyčejných nejmenších čtverců je i jednoduchost řešení úlohy (1.7), pro které existuje uzavřený tvar. K odvození tvaru řešení (1.7) využije následující dvě tvrzení. Řešení (1.7) lze odvodit i přímo, využitím Gramovy matice (Barto a Tůma (2019), Tvrzení 8.76). V kapitole 2 problém (1.7) zobecníme a bude již nutné využít níže uvedená tvrzení.

Lemma 1.3 (Subgradientní podmínka optimality, Lachout (2020), Lemma 2.55)**.** *Nechť* $\mathcal{D} \subseteq \mathbb{R}^p$, $\boldsymbol{x}^\star \in \mathcal{D}$ *a* $f : \mathcal{D} \to \mathbb{R}$ *je funkce. Poté* $\boldsymbol{x}^\star$ *je globálním minimem funkce* f *na množině* $\mathcal{D}$ *právě tehdy, když*

$$\mathbf{0} \in \partial f(\boldsymbol{x}^\star) = \{\boldsymbol{d} \in \mathbb{R}^p \,|\, \forall \boldsymbol{y} \in \mathcal{D} : f(\boldsymbol{y}) \geq \boldsymbol{d}^\top(\boldsymbol{y} - \boldsymbol{x}^\star) + f(\boldsymbol{x}^\star)\}. \tag{1.8}$$

Množina $\partial f(\boldsymbol{x})$ se nazývá *subdiferenciál* funkce f v bodě $\boldsymbol{x}$ a prvky množiny $\partial f(\boldsymbol{x})$ se nazývají *subgradienty.* Subgradietní podmínka optimality je speciálním případem *Karushových-Kuhnových-Tuckerových podmínek optimality* pro optimalizační úlohu bez vazebních podmínek. K nalezení subdiferenciálu účelové funkce úlohy (2.1) využijeme následující lemma.

Lemma 1.4 (Lachout (2020), lemma 2.56)**.** *Nechť* $\mathcal{G} \subseteq \mathbb{R}^p$ *je neprázdná konvexní množina,* $f : \mathcal{G} \to \mathbb{R}$ *je konvexní funkce a* $\boldsymbol{y} \in \mathcal{G}$*. Má-li funkce* f *gradient v bodě* $\boldsymbol{y}$*, pak* $\partial f(\boldsymbol{y}) = \{\nabla f(\boldsymbol{y})\}$.

Účelová funkce $\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$ je konvexní (plyne z lemmat A.1.1 a A.1.2), diferencovatelná na svém definičním oboru a v úloze (1.7) nejsou kladeny žádné vazební podmínky. Z lemmat 1.3 a 1.4 vyplývá, že řešením úlohy (1.7) je množina stacionárních bodů účelové funkce

$$\left\{\boldsymbol{\beta} \in \mathbb{R}^p : (\mathbb{X}^\top\mathbb{X})\boldsymbol{\beta} = \mathbb{X}^\top\boldsymbol{Y}\right\}. \tag{1.9}$$

Pozorování. Nechť $\mathbb{A} \in \mathbb{R}^{n\times p}$, $\boldsymbol{x} \in \mathbb{R}^p$ a $\boldsymbol{b} \in \mathbb{R}^n$. Označme $\mathbb{A}^+ \in \mathbb{R}^{p\times n}$ Mooreovu-Penroseovu pseudoinversi matice $\mathbb{A}$. Jestliže $\mathbb{A}\mathbb{A}^+\boldsymbol{b} = \boldsymbol{b}$, pak pro každé $\boldsymbol{\gamma} \in \mathbb{R}^p$ platí $\mathbb{A}(\mathbb{A}^+\boldsymbol{b} + (\mathbb{I}_n - \mathbb{A}^+\mathbb{A})\boldsymbol{\gamma}) = \boldsymbol{b}$ a všechna řešení soustavy lineárních rovnic $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ lze vyjádřit jako $\boldsymbol{x} = \mathbb{A}^+\boldsymbol{b} + (\mathbb{I}_n - \mathbb{A}^+\mathbb{A})\boldsymbol{\gamma}$. Navíc $(\mathbb{I}_n - \mathbb{A}^+\mathbb{A})$ je maticí ortogonální projekce na lineární prostor $\mathrm{Ker}(\mathbb{A}) = \{\boldsymbol{\gamma} \in \mathbb{R}^p : \mathbb{A}\boldsymbol{\gamma} = \boldsymbol{0}_n\}$, platí $\mathrm{Ker}(\mathbb{A}) = \{(\mathbb{I}_n - \mathbb{A}^+\mathbb{A})\boldsymbol{\gamma}, \boldsymbol{\gamma} \in \mathbb{R}^p\}$ a $\mathbb{A}^+\boldsymbol{b}$ je řešení soustavy $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ s nejmenší euklidovskou normou (minimalitu euklidovské normy řešení $\mathbb{A}^+\boldsymbol{b}$ lze získat jako důsledek Barto a Tůma (2019), Tvrzení 8.96).

Pro soustavu $(\mathbb{X}^\top\mathbb{X})\boldsymbol{\beta} = \mathbb{X}^\top\boldsymbol{Y}$ je splněno

$$(\mathbb{X}^\top\mathbb{X})(\mathbb{X}^\top\mathbb{X})^+\mathbb{X}^\top\boldsymbol{Y} = \mathbb{X}^\top\mathbb{X}\mathbb{X}^+\boldsymbol{Y} = \mathbb{X}^\top\boldsymbol{Y},$$

kde jsme postupně využili vztahů $(\mathbb{X}^\top\mathbb{X})^+\mathbb{X}^\top = \mathbb{X}^+$ a $\mathbb{X}^\top\mathbb{X}\mathbb{X}^+ = \mathbb{X}^\top$. Množina řešení soustavy rovnic (1.9) je tedy neprázdná a tvaru

$$\left\{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta} = (\mathbb{X}^\top\mathbb{X})^+\mathbb{X}^\top\boldsymbol{Y} + (\mathbb{I}_p - (\mathbb{X}^\top\mathbb{X})^+\mathbb{X}^\top\mathbb{X})\boldsymbol{\gamma}, \boldsymbol{\gamma} \in \mathbb{R}^p\right\},$$

což lze vyjádřit jako

$$\left\{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta} = \mathbb{X}^+\boldsymbol{Y} + \boldsymbol{\gamma}, \boldsymbol{\gamma} \in \mathrm{Ker}(\mathbb{X})\right\}. \tag{1.10}$$

Za předpokladu plné sloupcové hodnosti matice modelu $\mathbb{X}$, tedy $\mathrm{rank}(\mathbb{X}) = p$, je matice $\mathbb{X}^\top\mathbb{X}$ regulární (Barto a Tůma (2019), Tvrzení 8.80) a existuje právě jedno řešení problému obyčejných nejmenších čtverců (1.7), které je tvaru $(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y}$ (též za předpokladu plné sloupcové hodnosti matice $\mathbb{X}$ platí $\mathbb{X}^+ = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top$ a $\mathrm{Ker}(\mathbb{X}) = \{\boldsymbol{0}_p\}$). Odhad vektoru regresních koeficientů metodou obyčejných nejmenších čtverců je v případě $\mathrm{rank}(\mathbb{X}) = p$ nestranný

$$\mathbb{E}\widehat{\boldsymbol{\beta}}^{OLS} = \mathbb{E}(\mathbb{E}[\widehat{\boldsymbol{\beta}}^{OLS} \mid \mathbb{X}]) = \mathbb{E}((\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{E}[\boldsymbol{Y} \mid \mathbb{X}]) = \mathbb{E}((\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{X}\boldsymbol{\beta}^*) = \boldsymbol{\beta}^*$$

a podmíněný rozptyl odhadu je tvaru

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}^{OLS} \mid \mathbb{X}) = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathrm{Var}(\boldsymbol{Y} \mid \mathbb{X})\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1}.$$

Věta 1.5 (Asymptotické vlastnosti – Komárek (2021), Theorem 16.2 a Theorem 16.4)**.** *Nechť $n, p \in \mathbb{N}$, $(Y_i, X_{i,1}, \ldots, X_{i,p})^\top$, $i = 1, \ldots, n$, je náhodný výběr z rozdělení náhodného vektoru $(Y, \boldsymbol{X}^\top)^\top$, který splňuje lineární regresní model (1.2). Nechť existují smíšené momenty druhého řádu $\mathbb{E}|X_jX_k|$, $j, k = 1, \ldots, p$, a matice $\mathbb{W} = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)$ je pozitivně definitní. Poté je odhad vektoru regresních koeficientů metodou obyčejných nejmenších čtverců silně konzistentní*

$$\widehat{\boldsymbol{\beta}}^{OLS} \xrightarrow{s.j.} \boldsymbol{\beta}^*, \qquad n \to \infty.$$

Jestliže navíc pro všechna $j, k \in \{1, \ldots, p\}$ platí $\mathbb{E}|\epsilon^2X_jX_k| < \infty$, odhad metodou obyčejných nejmenších čtverců je i asymptoticky normální

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\boldsymbol{0}_p, \sigma^2\mathbb{W}^{-1}), \qquad n \to \infty. \tag{1.11}$$

Odhad metodou obyčejných nejmenších čtverců $\widehat{\boldsymbol{\beta}}^{OLS}$ je nejpopulárnějším odhadem vektoru regresních koeficientů v lineárním regresním modelu. Existují ale situace, kdy tento odhad není nejvhodnějším. Například obsahuje-li lineární regresní model nadbytečný počet vysvětlujících proměnných, predikce založené na odhadu $\widehat{\boldsymbol{\beta}}^{OLS}$ nemusí být spolehlivé. Přítomnost nadbytečných vysvětlujících proměnných je chybou modelu, se kterou si odhad $\widehat{\boldsymbol{\beta}}^{OLS}$ sám neporadí. V následující sekci 1.3 jsou uvedeny metody pro výběr podmnožiny vysvětlujících proměnných.

1.3 Výběr rysů

Jedním z hlavních úkolů statistické analýzy je výběr statistického modelu. V této práci se zabýváme pouze lineárním regresním modelem (1.2) a výběrem rysů, čili výběrem podmnožiny vysvětlujících proměnných, které mají signifikantní efekt na hodnotu odezvy. Pro množinu $\mathcal{M} \subseteq \{1, \ldots, p\}$ označme

$$\mathbb{X}_{\mathcal{M}} = (\boldsymbol{X}_j)_{j\in\mathcal{M}}, \quad \boldsymbol{X}_{\mathcal{M}} = (X_j)_{j\in\mathcal{M}}, \quad \boldsymbol{\beta}_{\mathcal{M}} = (\beta_j)_{j\in\mathcal{M}}, \quad \epsilon_{\mathcal{M}} = Y - \mathbb{E}(Y \mid \boldsymbol{X}_{\mathcal{M}}).$$

Množinu $\mathcal{M}$ budeme považovat za identifikátor lineárního regresního modelu

$$\mathfrak{M}_{\mathcal{M}} : Y = \boldsymbol{X}_{\mathcal{M}}^\top \boldsymbol{\beta}^{\mathcal{M}} + \epsilon_{\mathcal{M}}$$

a „*modelem* $\mathcal{M}$“ budeme označovat model $\mathfrak{M}_{\mathcal{M}}$. Vektor regresních koeficientů v modelu $\mathcal{M}$ značíme $\boldsymbol{\beta}^{\mathcal{M}}$ a jeho odhad budeme značit s vlnovkou $\widetilde{\boldsymbol{\beta}}_{\mathcal{M}}$. Tímto odlišujeme od značení $\widehat{\boldsymbol{\beta}}_{\mathcal{M}} = (\widehat{\beta}_j)_{j\in\mathcal{M}}$, kde $\widehat{\boldsymbol{\beta}}$ je odhad $\boldsymbol{\beta}^*$ v plném modelu. V modelu $\mathcal{M}$ není pro $j \notin \mathcal{M}$ dostupná informace o vysvětlující proměnné X_j. Efekt X_j na odezvu Y, který není obsažený v $\boldsymbol{X}_{\mathcal{M}}$, je schován v náhodné chybě $\epsilon_{\mathcal{M}}$.

1.3.1 Trénovací a testovací chyba

Nechť $\mathcal{M}_0 \subset \mathcal{M}_1 \subseteq \{1, \ldots, p\}$. Model $\mathcal{M}_0$ budeme nazývat *podmodelem* modelu $\mathcal{M}_1$. Model $\mathcal{M}_1$ neobsahuje méně informace, než jeho libovolný podmodel $\mathcal{M}_0$. Alternativně $\sigma(\boldsymbol{X}_{\mathcal{M}_0}) \subseteq \sigma(\boldsymbol{X}_{\mathcal{M}_1})$ a reziduální rozptyl v modelu $\mathcal{M}_1$ je menší nebo roven reziduálnímu rozptylu v podmodelu $\mathcal{M}_0$

$$\sigma^2_{\mathcal{M}_1} = \mathbb{E}(Y - \mathbb{E}(Y \mid \boldsymbol{X}_{\mathcal{M}_1}))^2 \leq \mathbb{E}(Y - \mathbb{E}(Y \mid \boldsymbol{X}_{\mathcal{M}_0}))^2 = \sigma^2_{\mathcal{M}_0}. \tag{1.12}$$

Může se proto zdát, že model obsahující více vysvětlujících proměnných bude vždy lepší, neboť dosahuje nižšího reziduálního rozptylu. Tato úvaha ale není správná pro odhady sestrojené na základě omezeného počtu pozorování n. Uvažujme průměr čtvercové chyby predikce na *trénovacích datech* (trénovací data jsou data, která byla použita k sestrojení odhadu), kterou budeme nazývat *trénovací chyba*

$$\frac{1}{n}\|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|_2^2$$

a průměr čtvercové chyby predikce na *testovacích datech* (testovací data jsou dostupná data $(Y_i, \boldsymbol{X}_i^\top)^\top$, $i = n+1, \ldots, n+m$, $m \in \mathbb{N}$, která nebyla použita k sestrojení odhadu), kterou budeme nazývat *testovací chyba*

$$\frac{1}{m}\sum_{i=n+1}^{n+m}(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})^2, \tag{1.13}$$

Označme $\boldsymbol{P}_{\mathbb{X}} = \mathbb{X}\mathbb{X}^+$ *matici ortogonální projekce* na lineární prostor $\mathrm{Im}(\mathbb{X})$. Předpokládejme plnou sloupcovou hodnost matice modelu $\mathbb{X}$. Střední hodnota tréninkové chyby odhadu metodou obyčejných nejmenších čtverců je

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}\left[\frac{1}{n}\|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{OLS}\|_2^2 \mid \mathbb{X}\right]\right) &= \frac{1}{n}\mathbb{E}(\mathbb{E}[\|(\mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}) - \boldsymbol{P}_{\mathbb{X}}(\mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon})\|_2^2 \mid \mathbb{X}]) \\
&= \frac{1}{n}\mathbb{E}(\mathbb{E}[\|\boldsymbol{\epsilon} - \boldsymbol{P}_{\mathbb{X}}\boldsymbol{\epsilon}\|_2^2 \mid \mathbb{X}]) \\
&= \frac{1}{n}\mathbb{E}\left(\|\boldsymbol{\epsilon}\|_2^2 - 2\mathbb{E}[\boldsymbol{\epsilon}^\top \boldsymbol{P}_{\mathbb{X}}\boldsymbol{\epsilon} \mid \mathbb{X}] + \mathbb{E}[\|\boldsymbol{P}_{\mathbb{X}}\boldsymbol{\epsilon}\|_2^2 \mid \mathbb{X}]\right) \\
&= \frac{n\sigma^2 - 2p\sigma^2 + p\sigma^2}{n} = \frac{n-p}{n}\sigma^2.
\end{aligned}$$

Označme

$$\boldsymbol{Y}_{test} = \begin{pmatrix} Y_{n+1} \\ \vdots \\ Y_{n+m} \end{pmatrix}, \quad \mathbb{X}_{test} = \begin{pmatrix} X_{n+1,1} & \dots & X_{n+1,p} \\ \vdots & \ddots & \vdots \\ X_{n+m,1} & \dots & X_{n+m,p} \end{pmatrix}, \quad \boldsymbol{\epsilon}_{test} = \begin{pmatrix} \epsilon_{n+1} \\ \vdots \\ \epsilon_{n+m} \end{pmatrix}.$$

K vyjádření testovací chyby odhadu metodou obyčejných nejmenších čtverců v lineárním regresním modelu nejprve vyjádříme podmíněnou střední hodnotu

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{Y}_{test} - \mathbb{X}_{test}\widehat{\boldsymbol{\beta}}^{OLS}\|_2^2 \,|\, \mathbb{X},\, \mathbb{X}_{test}\right] = \\
&= \mathbb{E}[\|(\boldsymbol{Y}_{test} - \mathbb{X}_{test}\boldsymbol{\beta}^*) - (\mathbb{X}_{test}\widehat{\boldsymbol{\beta}}^{OLS} - \mathbb{X}_{test}\boldsymbol{\beta}^*)\|_2^2 \,|\, \mathbb{X},\, \mathbb{X}_{test}] \\
&= \mathbb{E}\|\boldsymbol{\epsilon}_{test}\|_2^2 + \mathbb{E}[\|\mathbb{X}_{test}\widehat{\boldsymbol{\beta}}^{OLS} - \mathbb{X}_{test}\boldsymbol{\beta}^*\|_2^2 \,|\, \mathbb{X},\, \mathbb{X}_{test}] \\
&- 2\sum_{i=n+1}^{n+m}\left(\mathbb{E}(\epsilon_i)\mathbb{E}[\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{X}_i^\top\boldsymbol{\beta}^* \,|\, \mathbb{X},\, \mathbb{X}_{test}]\right) \\
&= \mathbb{E}\|\boldsymbol{\epsilon}_{test}\|_2^2 + \mathbb{E}[\|\mathbb{X}_{test}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top(\mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}) - \mathbb{X}_{test}\boldsymbol{\beta}^*\|_2^2 \,|\, \mathbb{X},\, \mathbb{X}_{test}] \\
&= m\sigma^2 + \mathbb{E}[\|\mathbb{X}_{test}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{\epsilon}\|_2^2 \,|\, \mathbb{X},\, \mathbb{X}_{test}].
\end{aligned}$$

Za předpokladu $\mathbb{X}_{test} = \mathbb{X}$ hovoříme o *predikci replikované odezvy*, platí

$$\mathbb{E}[\|\mathbb{X}_{test}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{\epsilon}\|_2^2 \,|\, \mathbb{X} = \mathbb{X}_{test}] = \mathbb{E}[\|\boldsymbol{P}_{\mathbb{X}}\boldsymbol{\epsilon}\|_2^2 \,|\, \mathbb{X}] = \operatorname{tr}(\operatorname{Var}(\boldsymbol{P}_{\mathbb{X}}\boldsymbol{\epsilon} \,|\, \mathbb{X})) = p\sigma^2$$

a střední hodnota testovací chyby je $\mathbb{E}[\mathbb{E}[\|\boldsymbol{Y}_{test} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{OLS}\|_2^2/n \,|\, \mathbb{X}]] = \sigma^2(n+p)/n$. Využitím Groves a Rothenberg (1969) lze dokázat, že za předpokladu pozitivně definitní matice $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)$ je střední hodnota testovací chyby vždy větší nebo rovna střední hodnotě testovací chyby pro replikovanou odezvu. Tedy zatímco s fixním počtem pozorování $n \in \mathbb{N}$ a rostoucím počtem parametrů modelu p tréninková chyba klesá, testovací chyba klesat nemusí.

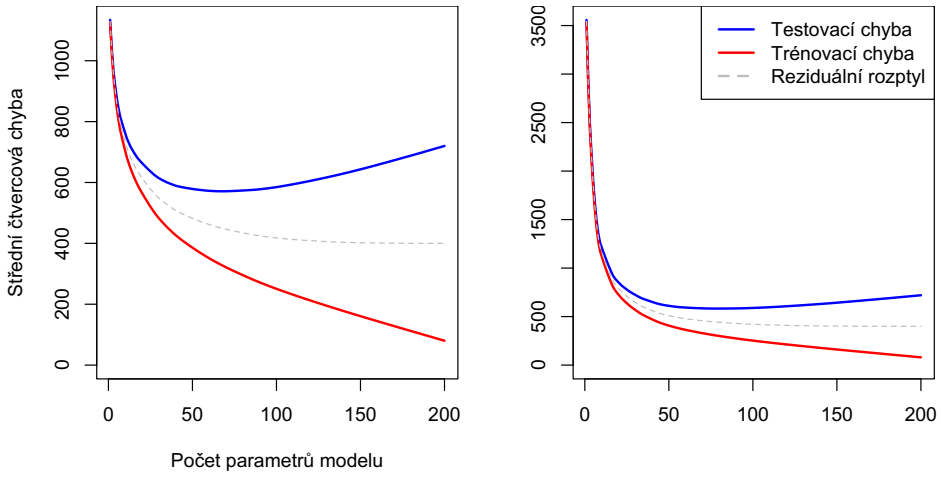
Příklad. Nechť $p \in \mathbb{N}$, $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{0}_p, \sigma_X^2\mathbb{I}_p)$, $\sigma_X^2 > 0$ a $\zeta \sim \mathcal{N}(0,\, \sigma_\zeta^2)$ je náhodná veličina nezávislá s $\boldsymbol{X}$. Nechť $\eta_1, \dots, \eta_p \in \mathbb{R}$ je náhodný výběr ze směsi dvou centrovaných Laplaceových rozdělení s hustotou

$$\pi(\eta_j; \gamma_j,\, \lambda_1,\, \lambda_2) = \gamma_j \cdot \frac{\lambda_1}{2}e^{-\lambda_1|\eta_j|} + (1-\gamma_j)\cdot\frac{\lambda_2}{2}e^{-\lambda_2|\eta_j|}, \qquad j = 1, \dots, p,$$

kde $\gamma_j \in \{0,\, 1\}$ je parametr určující rozdělení η_j a $\lambda_1,\, \lambda_2 \in \mathbb{R}$, $\lambda_1 > \lambda_2 > 0$ jsou parametry Laplaceova rozdělení. Rozdělení s parametrem λ_1 je „*špičaté*“ a rozdělení s parametrem je λ_2 „*placaté*“. Regresní koeficienty vygenerované ze špičatého, resp. placatého rozdělení typicky přísluší vysvětlujícím proměnným, které mají malý, resp. velký vliv na hodnotu odezvy. Označme $\eta_{(j)}$ j-tou pořádkovou statistiku z náhodného výběru $\eta_1, \dots, \eta_p$ a $\beta_1^\star = \eta_{(p)}, \dots, \beta_p^\star = \eta_{(1)}$, čili platí $\beta_1^\star \geq \dots \geq \beta_p^\star$. Uvažujme odezvu $Y = \boldsymbol{X}^\top\boldsymbol{\beta}^\star + \zeta$ a lineární regresní model

$$Y = \sum_{j=1}^{q}\beta_j^\star X_j + \epsilon_q, \tag{1.14}$$

kde $q \in \mathbb{N}$, $q \leq p$ je počet parametrů modelu a $\epsilon_q = \zeta + \sum_{j=q+1}^{p}\beta_j^\star X_j$ je náhodná chyba v modelu (1.14). Zřejmě $\epsilon_q \sim \mathcal{N}(0,\, \sigma_q^2)$, kde $\sigma_q^2 = \sigma_\zeta^2 + \sum_{j=q+1}^{p}(\beta_j^\star)^2\sigma_X^2$. Na obrázku 1.1 níže je vykreslena střední hodnota testovací a trénovací chyby odhadu metodou obyčejných nejmenších čtverců v modelu (1.14) jako funkce q.



(a) $\lambda_1 = 1$, $\lambda_2 = 0{,}25$ (b) $\lambda_1 = 1$, $\lambda_2 = 0{,}1$

Obrázek 1.1: Střední hodnota trénovací chyby a testovací chyby pro replikovanou odezvu odhadu metodou nejmenších čtverců v modelu (1.14), jako funkce počtu parametrů q. Rozsah výběru $n = 250$, počet pozorovaných parametrů $p = 200$, reziduální rozptyl v plném modelu $\sigma_\zeta^2 = 400$, rozptyl vysvětlujících proměnných $\sigma_X^2 = 1$. Padesát regresních koeficientů bylo vygenerováno z Laplaceova rozdělení s parametrem λ_2, zbylé koeficienty byly vygenerovány z Laplaceova rozdělení s parametrem λ_1. Vygenerované regresní koeficienty byly následně seřazeny sestupně.

Odhad střední hodnoty testovací chyby se běžně provádí pomocí *křížové validace*. Při křížové validaci jsou data rozdělena do $K \in \mathbb{N}$ stejně velkých podmnožin. Jako trénovací data jsou použita data pouze z $K - 1$ podmnožin a jako testovací data jsou použita data ze zbylé podmnožiny. Na testovacích datech lze následně prováděny různé analýzy, jako například výpočet testovací chyby. Celý tento proces je opakován K-krát, přičemž během každého opakování je zvolena jiná z K podmnožin dat za testovací data. Průměr z vypočtených testovacích chyb se nazývá *chyba křížové validace*. Chyba křížové validace se používá jako odhad střední hodnoty testovací chyby a jako míra kvality predikcí modelu. Jedná se o flexibilní metriku, kterou lze použít v široké škále statistických modelů.

Jev, kdy je tréninková chyba nízká, ale testovací chyba vysoká, se nazývá *přetrénování*. Predikce na základě přetrénovaného modelu zřejmě nemohou být spolehlivé. K přetrénování dochází zejména nemáme-li dostatečný počet pozorování na počet parametrů zvoleného podmodelu. V takovém případě odhad vytváří asociace, které fungují na pozorovaných datech, ale obecně nedávají smysl.

Minimální vhodný počet pozorování závisí na mnoha faktorech, jako je například zvolený model, účel analýzy, v případě lineární regrese *poměru signálu a šumu* $\|\boldsymbol{\beta}^*\|_2^2/\sigma^2$ a v případě testování vlivu regresních koeficientů zvolený test a síla testu (typicky $1 - \beta = 1 - 4 \cdot \alpha = 0{,}8$). Výpočet dostatečného počtu pozorování pro danou analýzu nebývá jednoduchý a tak lze v literatuře nalézt různá pravidla palce pro minimální počet pozorování. Běžně se doporučený minimální počet pozorování na jednu vysvětlující proměnnou pohybuje mezi pěti až dvaceti. Překvapivý výsledek lze nalézt v Austin a Steyerberg (2015), kde pro predikce odhadu metodou obyčejných nejmenších čtverců v lineárním regresním modelu uvádí za dostatečné již dvě pozorování na jednu vysvětlující proměnnou. Upozorněme, že v tomto článku pracují s reálnými daty, ale v následných simulacích je skutečná odezva nahrazena odezvou nasimulovanou z lineárního regresního modelu $Y \mid \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^{OLS}, \widehat{\sigma}^2)$ (uvedeno v sekci 3.2.), kde $\widehat{\boldsymbol{\beta}}^{OLS}$ a $\widehat{\sigma}^2$ jsou odhady na základě dostupných pozorování. Hodnotu $\widehat{\sigma}^2$ jsme v článku nenalezli.

1.3.2 Testování podmodelů

Pro odhad vektoru regresních koeficientů metodou obyčejných nejmenších čtverců lze odvodit různá asymptotická rozdělení, pomocí kterých lze testovat signifikanci podmnožiny složek. Test signifikance podmnožiny složek vektoru regresních koeficientů je základní metodou výběru rysů – do modelu zvolíme pouze vysvětlující proměnné, u kterých zamítneme hypotézu nulového efektu.

Dle Komárek (2021), Theorem 16.3

$$\widehat{\sigma}^2 = \frac{\|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{OLS}\|_2^2}{n-p} \xrightarrow{s.j.} \sigma^2, \qquad n \to \infty.$$

Využitím centrální limitní věty na jednotlivé složky matice $\mathbb{X}^\top\mathbb{X}$

$$\widehat{\mathbb{W}}_n = \frac{1}{n}\mathbb{X}^\top\mathbb{X} \xrightarrow{s.j.} \mathbb{W}, \qquad n \to \infty.$$

Po vzoru Fu a Knight (2000) budeme předpokládat, že matice $\mathbb{W}$ je pozitivně definitní. Z asymptotické normality odhadu metodou obyčejných nejmenších čtverců společně s delta metodou a Cramérovou–Slutského větou

$$\forall \boldsymbol{l} \in \mathbb{R}^p, \boldsymbol{l} \neq \boldsymbol{0}_p : \frac{\sqrt{n}(\boldsymbol{l}^\top\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{l}^\top\boldsymbol{\beta})}{\sqrt{\boldsymbol{l}^\top(\widehat{\sigma}^2\widehat{\mathbb{W}}_n^{-1})\boldsymbol{l}}} \xrightarrow{s.j.} \mathcal{N}(0,\,1), \qquad n \to \infty. \tag{1.15}$$

Asymptotického rozdělení (1.15) lze využít k testu nulové hypotézy $H_0 : \beta_j^* = 0$ vůči alternativě $H_1 : \beta_j^* \neq 0$. Nulovou hypotézu zamítáme v případě

$$\frac{|\widehat{\beta}_j^{OLS}|}{\sqrt{\widehat{\sigma}^2\widehat{v}_{j,j}}} \geq \Phi^{-1}(1-\alpha/2),$$

kde $\widehat{v}_{j,j} = ([\mathbb{X}^\top\mathbb{X}]^{-1})_{j,j}$, $\Phi^{-1}(1-\alpha/2)$ je $1-\alpha/2$ kvantil normovaného normálního rozdělení $\mathcal{N}(0,\,1)$ a α je zvolená hladina testu, typicky $\alpha = 0{,}05$.

Často je vhodné testovat signifikanci více regresních koeficientů naráz. Uvažujme předem zvolený pár model-podmodel $\mathcal{M}_0 \subset \mathcal{M}_1$. K testování hypotézy $H_0 : \forall j \in (\mathcal{M}_1 \setminus \mathcal{M}_0) : \beta_j^{\mathcal{M}_1} = 0$ (podmodel $\mathcal{M}_0$ neobsahuje signifikantně méně informace, než model $\mathcal{M}_1$) vůči alternativě $H_1 : \exists j \in (\mathcal{M}_1 \setminus \mathcal{M}_0) : \beta_j^{\mathcal{M}_1} \neq 0$ (model $\mathcal{M}_1$ obsahuje signifikantní informaci, kterou podmodel $\mathcal{M}_0$ neobsahuje) se používá testová statistika

$$F_{\mathcal{M}_0,\mathcal{M}_1} = \frac{\|\boldsymbol{Y} - \mathbb{X}_{\mathcal{M}_0}\widetilde{\boldsymbol{\beta}}_{\mathcal{M}_0}^{OLS}\|_2^2 - \|\boldsymbol{Y} - \mathbb{X}_{\mathcal{M}_1}\widetilde{\boldsymbol{\beta}}_{\mathcal{M}_1}^{OLS}\|_2^2}{(p_{\mathcal{M}_1} - p_{\mathcal{M}_0})\widehat{\sigma}_{\mathcal{M}_1}^2}, \tag{1.16}$$

kde $p_{\mathcal{M}} = |\mathcal{M}|$ a

$$\widehat{\sigma}_{\mathcal{M}_1}^2 = \frac{\|\boldsymbol{Y} - \mathbb{X}_{\mathcal{M}_1}\widetilde{\boldsymbol{\beta}}_{\mathcal{M}_1}^{OLS}\|_2^2}{n - p_{\mathcal{M}_1}} \tag{1.17}$$

je nestranný konzistentní odhad reziduálního rozptylu $\sigma_{\mathcal{M}_1}^2$ v modelu $\mathcal{M}_1$. Za platnosti nulové hypotézy H_0 má dle Komárek (2021), Theorem 16.4, testová statistika $F_{\mathcal{M}_0,\mathcal{M}_1}$ asymptoticky χ^2 rozdělení o $p_{\mathcal{M}_1} - p_{\mathcal{M}_0}$ stupních volnosti

$$F_{\mathcal{M}_0,\mathcal{M}_1} \xrightarrow{\mathcal{D}} \chi^2_{p_{\mathcal{M}_1} - p_{\mathcal{M}_0}}, \qquad n \to \infty.$$

Za předpokladu platnosti nulové hypotézy a normálního lineárního regresního modelu má testová statistika $F_{\mathcal{M}_0,\mathcal{M}_1}$ rozdělení F o $p_1 - p_0$ a $n - p$ stupních volnosti (Komárek (2021), Theorem 8.1).

1.3.3 Riziko odhadu a střední čtvercová chyba

Nechť $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ je náhodný výběr z p-rozměrného rozdělení s distribuční funkcí $F_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\theta})$, která závisí na neznámém parametru $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^q$, $q \in \mathbb{N}$. Pro odhad $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ parametru $\boldsymbol{\theta}$ definujeme *rizikovou funkci odhadu*

$$R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})) = \int_{\{(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \,:\, \boldsymbol{z}_i \in \mathbb{R}^p,\, i=1,\ldots,n\}} L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) dF_{\boldsymbol{Z}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n; \boldsymbol{\theta}),$$

kde $\mathbb{E}_{\boldsymbol{\theta}}$ značí střední hodnotu vzhledem k rozdělení pozorovaných dat se sdruženou distribuční funkcí $F_{\boldsymbol{Z}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n; \boldsymbol{\theta})$ a $L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) : \boldsymbol{\Theta} \times \mathbb{R}^q \to \mathbb{R}$ je zvolená *ztrátová funkce*. Riziková funkce se používá jako míra kvality odhadu. Asi nejpopulárnější rizikovou funkcí je *střední čtvercová chyba odhadu*

$$\begin{aligned}
\mathrm{MSE}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2 &= (\|\boldsymbol{\theta}\|_2^2 - 2\boldsymbol{\theta}^\top \mathbb{E}_{\boldsymbol{\theta}} \widehat{\boldsymbol{\theta}} + \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}\|_2^2) + \|\mathbb{E}_{\boldsymbol{\theta}} \widehat{\boldsymbol{\theta}}\|_2^2 - \|\mathbb{E}_{\boldsymbol{\theta}} \widehat{\boldsymbol{\theta}}\|_2^2 \\
&= (\|\boldsymbol{\theta}\|_2^2 - 2\boldsymbol{\theta}^\top \mathbb{E}_{\boldsymbol{\theta}} \widehat{\boldsymbol{\theta}} + \|\mathbb{E}_{\boldsymbol{\theta}} \widehat{\boldsymbol{\theta}}\|_2^2) + \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}\|_2^2 - \|\mathbb{E}_{\boldsymbol{\theta}} \widehat{\boldsymbol{\theta}}\|_2^2 \\
&= \|\mathbb{E}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|_2^2 + \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}\|_2^2 - \|\mathbb{E}_{\boldsymbol{\theta}} \widehat{\boldsymbol{\theta}}\|_2^2 \\
&= \|\mathrm{Bias}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})\|_2^2 + \mathrm{tr}(\mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})),
\end{aligned}$$

kde

$$\mathrm{Bias}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

je *vychýlení odhadu* $\widehat{\boldsymbol{\theta}}$ při dané hodnotě parametru $\boldsymbol{\theta}$ a

$$\mathrm{tr}(\mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})) = \sum_{j=1}^{p} \mathrm{var}_{\boldsymbol{\theta}}(\widehat{\theta}_j) = \sum_{j=1}^{p} [\mathbb{E}_{\boldsymbol{\theta}}(\widehat{\theta}_j^{\,2}) - (\mathbb{E}_{\boldsymbol{\theta}} \widehat{\theta}_j)^2] = \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}\|_2^2 - \|\mathbb{E}_{\boldsymbol{\theta}} \widehat{\boldsymbol{\theta}}\|_2^2.$$

Střední čtvercová chyba odhadu podmíněné střední hodnoty $\widehat{\mathbb{E}[\boldsymbol{Y} \,|\, \mathbb{X}]} = \mathbb{X}\widehat{\boldsymbol{\beta}}$ v lineárním regresním modelu je úzce svázána s přesností predikce, například střední hodnotu tréninkové chyby lze (až na konstantu n^{-1}) vyjádřit jako

$$\mathbb{E} \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|_2^2 = \sigma^2 + \mathrm{MSE}(\mathbb{X}\boldsymbol{\beta}, \mathbb{X}\widehat{\boldsymbol{\beta}}).$$

Věta 1.6 (Gaussova–Markovova)**.** *Předpokládejme platnost lineárního regresního modelu a plnou sloupcovou hodnosti matice modelu. Odhad metodou obyčejných nejmenších čtverců dosahuje nejnižší střední čtvercové chyby podmíněně pozorovanými daty, mezi všemi nestrannými lineárními odhady. Tedy*

$$\widehat{\boldsymbol{\beta}}^{OLS} = \operatorname*{argmin}_{\widehat{\boldsymbol{\beta}}} \mathbb{E}_{\boldsymbol{\beta}}(\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2^2 \,|\, \mathbb{X}) = \operatorname*{argmin}_{\widehat{\boldsymbol{\beta}}} \mathrm{MSE}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} | \, \mathbb{X}),$$

kde $\widehat{\boldsymbol{\beta}}$ je nestranný lineární odhad $\boldsymbol{\beta}$. Střední čtvercová chyba tohoto odhadu je rovna

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}^{OLS}, \boldsymbol{\beta} | \, \mathbb{X}) = \mathrm{tr}(\mathrm{Var}(\widehat{\boldsymbol{\beta}}^{OLS} \,|\, \mathbb{X})) = \sigma^2 \sum_{j=1}^{p} \frac{1}{d_j},$$

kde $d_j > 0$, $j = 1, \ldots, p$ jsou vlastní čísla matice $\mathbb{X}^\top \mathbb{X}$.

Z Věty 1.6 plyne, že libovolný lineární odhad vektoru regresních koeficientů, jehož podmíněná střední čtvercová chyba je nižší, než podmíněná střední čtvercová chyba odhadu metodou obyčejných nejmenších čtverců, musí být vychýlený. Je-li možné dosáhnout velké redukce střední čtvercové chyby za cenu malého vychýlení, je na místě využití vychýleného odhadu. Argumentem k této úvaze může být i citát Geogre E. P. Boxe „*Všechny modely jsou chybné, ale některé jsou užitečné*“. Tedy i odhad $\widehat{\boldsymbol{\beta}}^{OLS}$ lze považovat za vychýlený a jiný, více vychýlený odhad s menší střední čtvercovou chybou, lze považovat za více užitečný.

1.4 Regularizace

V následujícím se budeme zabývat vyhodnocením efektu vysvětlujících proměnných na podmíněnou střední hodnotu odezvy $\mathbb{E}[Y \mid \boldsymbol{X}]$ v *řídkém lineárním regresním modelu*. Řídký lineární regresní model je lineární regresní model, ve kterém je počet parametrů p vysoký, ale počet nenulových parametrů $p_{\mathcal{S}} = \|\boldsymbol{\beta}^*\|_0$ je nízký. V důsledku vysokého počtu parametrů p hrozí $\operatorname{rank} \mathbb{X} \leq n \ll p$, přičemž by pro $\mathcal{S} = \{j : \beta_j^* \neq 0\}$ stále mohlo platit $\operatorname{rank} \mathbb{X}_{\mathcal{S}} = p_{\mathcal{S}} \leq n$. Dle (1.10) odhad vektoru regresních koeficientů metodou obyčejných nejmenších čtverců není v případě $n < p$ jednoznačný a navíc pro alespoň jedno $j \in \{1, \ldots, p\}$ existují odhady $\widehat{\beta}_j^{OLS} > 0$ a $\overline{\beta}_j^{OLS} < 0$, což znemožňuje interpretaci tohoto odhadu. I v případě, kdy matice modelu má plnou sloupcovou hodnost, nemusí být odhad metodou obyčejných nejmenších čtverců tím nejvhodnějším, neboť zejména jsou-li si hodnoty p a n blízké, může docházet k přetrénování nebo *multikolinearitě*[1]. Výše zmíněné problémy lze řešit pomocí *regularizace*.

Regularizací se v našem kontextu míní zavedení restrikcí na parametr $\boldsymbol{\beta}$ v problému (1.7), které vedou k „ztlumeným odhadům“. Složky regularizovaného odhadu jsou obvykle menší v absolutní hodnotě a obvykle mají menší rozptyl, než složky odhadu metodou obyčejných nejmenších čtverců. Zavedené restrikce lze formulovat ve *vázané formě* jako dodatečné podmínky na množinu přípustných řešení

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2, \qquad \text{s.t. } \boldsymbol{\beta} \in \mathcal{C}, \tag{1.18}$$

kde množina $\mathcal{C} \subseteq \mathbb{R}^p$ je typicky konvexní, nebo v *penalizované formě*, přičtením typicky konvexní *penalizační funkce* $g : \mathbb{R}^p \to \mathbb{R}$ k účelové funkci

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\beta}). \tag{1.19}$$

Na penalizační funkci g lze nahlížet jako na funkci, která kvantifikuje preferenci určité hypotézy oproti jiným hypotézám. Pro řídké regresní modely je preferovanou hypotézou $\|\boldsymbol{\beta}^*\|_0 \leq h(\lambda) < p$, kde $h : [0, \infty) \to \{p-1, p-2, \ldots, 1, 0\}$ je nerostoucí funkce a $\lambda \geq 0$ volíme.

Příklad (Regularizované odhady). Vázaná forma, $k \in \mathbb{N}$, $t \geq 0$,

1. $\min\limits_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$ s.t. $\|\boldsymbol{\beta}\|_0 \leq k$, (výběr podmnožiny)
2. $\min\limits_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$ s.t. $\|\boldsymbol{\beta}\|_1 \leq t$, (lasso)
3. $\min\limits_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$ s.t. $\|\boldsymbol{\beta}\|_2^2 \leq t$. (hřebenová regrese)

Penalizovaná forma, $\lambda \geq 0$ se nazývá *regularizační parametr*

4. $\widehat{\boldsymbol{\beta}}^{Subset}(\lambda) \in \operatorname*{argmin}\limits_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0$, (výběr podmnožiny)
5. $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda) \in \operatorname*{argmin}\limits_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$, (lasso)
6. $\widehat{\boldsymbol{\beta}}^{Ridge}(\lambda) \in \operatorname*{argmin}\limits_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$. (hřebenová regrese)

[1] Multikolinearita je jev, kdy jsou sloupce matice modelu téměř lineárně závislé.

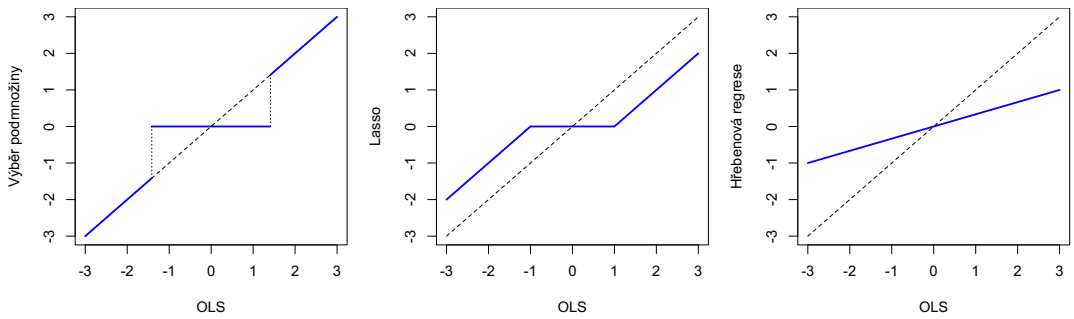
V této práci se budeme primárně věnovat problému lasso v penalizované formě. Pro přehlednost budeme příležitostně vynechávat λ v argumentu odhadu $\widehat{\boldsymbol{\beta}}(\lambda)$ a tento princip budeme využívat i u jiných veličin. Regularizační parametr λ je běžně vybírán expertně, nebo jako argument minima zvoleného kritéria. Často se například volí λ, které minimalizuje chybu křížové validace.

Poznámka. Rozšířením řídkých lineárních regresních modelů jsou modely, ve kterých je počet parametrů p vysoký, ale počet regresních koeficientů, které pro zvolenou hodnotu $\epsilon \geq 0$ splňují $|\beta_j^*| > \epsilon$ je nízký. Za *relevantní vysvětlující proměnné* jsou poté považovány vysvětlující proměnné $\{X_j : |\beta_j^*| > \epsilon\}$. Některé regularizované odhady, například výběr podmnožiny, odhadnou určité množství regresních koeficientů jako nulové. *Množinu aktivních složek* odhadu $\{j : \widehat{\beta}_j \neq 0\}$ lze poté považovat za odhad množiny relevantních vysvětlujících proměnných. Jaké proměnné jsou považovány za relevantní závisí u výběru podmnožiny na volbě regularizačního parametru λ. Při odhadu množiny relevantních vysvětlujících proměnných lze vychýlení odhadu irelevantních regresních koeficientů směrem k nule považovat za žádoucí vlastnost.

Výběr podmnožiny se z výše zmíněných metod může jevit jako nejlepší, neboť provádí odhad množiny relevantních vysvětlujících proměnných a v modelu, který obsahuje pouze vybrané vysvětlující proměnné produkuje nestranné odhady. Praktické využití této metody je ale limitováno, neboť se jedná o problém celočíselného programování, který je výpočetně náročný. Lasso a hřebenová regrese jsou problémy konvexního programování a jsou snadno řešitelné. Brzy navíc ukážeme, že i metoda lasso provádí odhad množiny relevantních vysvětlujících proměnných. Hřebenová regrese neprovádí odhad množiny relevantních vysvětlujících proměnných a nejedná se o vhodnou metodu pro řídké regresní modely. Díky konvexitě je u metod lasso a hřebenová regrese splněna silná dualita a pro každé λ z penalizované formy existuje t z vázané formy takové, že problémy ve vázané a penalizované formě mají stejně řešení. Pro výběr podmnožiny vázaná a penalizovaná forma nejsou ekvivalentní. V případě ortogonální matice modelu

$$\widehat{\boldsymbol{\beta}}^{Subset}(\lambda) = H_{\sqrt{2\lambda}}(\widehat{\boldsymbol{\beta}}^{OLS}), \qquad \widehat{\boldsymbol{\beta}}^{Lasso}(\lambda) = S_\lambda(\widehat{\boldsymbol{\beta}}^{OLS}), \qquad \widehat{\boldsymbol{\beta}}^{Ridge}(\lambda) = \frac{\widehat{\boldsymbol{\beta}}^{OLS}}{1+2\lambda},$$

kde $H_\lambda(\boldsymbol{x}) = \boldsymbol{x} \odot \mathbb{I}[|\boldsymbol{x}| \geq \lambda]$ a $S_\lambda(\boldsymbol{x}) = \text{sign}(\boldsymbol{x}) \odot (|\boldsymbol{x}| - \lambda)_+$ (součin $\odot$ a aplikované funkce míněny po složkách) jsou *tvrdá* a *měkká prahová funkce* a $x_+ = \max\{0, x\}$.



(a) Výběr podmnožiny (b) Metoda lasso (c) Hřebenová regrese

Obrázek 1.2: Odhady regresního koeficientu v případě ortogonální matice modelu metodami výběr podmnožiny, lasso a hřebenová regrese s volbou regularizačního parametru $\lambda = 1$, jako funkce odhadu metodou obyčejných nejmenších čtverců (OLS).

Prahování slouží k výběru relevantních vysvětlujících proměnných a ideálně vede ke konzistentnímu odhadu množiny aktivních složek vektoru regresních koeficientů $\mathcal{S} = \mathrm{Supp}(\boldsymbol{\beta}^*) = \{j : \beta_j^* \neq 0\}$. U odhadů $\widehat{\boldsymbol{\beta}}^{Lasso}$ a $\widehat{\boldsymbol{\beta}}^{Ridge}$ dále dochází k *smrštění*[2], tedy vychýlení odhadu směrem k nule. Prahování i smrštění mají vést k redukci variability[3] odhadu za cenu vyššího vychýlení. Z metod výběr podmnožiny, lasso a hřebenová regrese existuje obecné řešení v uzavřeném tvaru pouze pro hřebenovou regresi

$$\widehat{\boldsymbol{\beta}}^{Ridge}(\lambda) = (\mathbb{X}^\top\mathbb{X} + 2\lambda\mathbb{I}_p)^{-1}\mathbb{X}^\top\boldsymbol{Y}. \tag{1.20}$$

Výhodou výše zmíněných regularizovaných odhadů je, že na rozdíl od metody obyčejných nejmenších čtverců nevyžadují k jednoznačnosti řešení plnou sloupcovou hodnost matice modelu. U hřebenové regrese přičtení penalizační funkce vede k „regularizaci“ invertované matice v (1.20). U výběru podmnožiny a metody lasso dochází k „regularizaci“ díky prahování, které vede k výběru množiny aktivních složek odhadu $\mathcal{A}(\lambda) = \mathrm{Supp}(\widehat{\boldsymbol{\beta}}(\lambda))$ tak, aby matice $\mathbb{X}_{\mathcal{A}(\lambda)}$ měla plnou sloupcovou hodnost. V případě spojitých centrovaných vysvětlujících proměnných je ovšem dále nutné $|\mathcal{A}(\lambda)| \leq n - 1$, což nemusí být splněno pro všechna $\lambda \in [0, \infty)$.

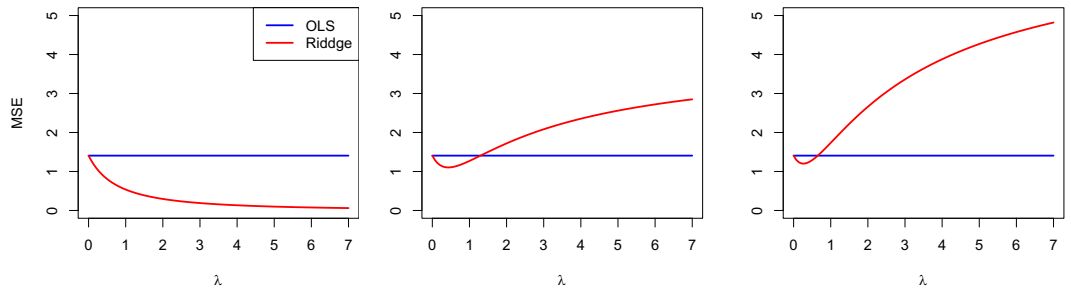
Příklad (Kompromis mezi rozptylem a vychýlením). Označme d_j, $j = 1, \ldots, p$, vlastní čísla matice $\mathbb{X}^\top\mathbb{X}$. Za předpokladu plné sloupcové hodnosti matice modelu

$$\mathrm{tr}(\mathrm{Var}(\widehat{\boldsymbol{\beta}}^{Ridge} \,|\, \mathbb{X})) = \sigma^2 \sum_{j=1}^{p} \frac{d_j}{(d_j + 2\lambda)^2} \leq \sigma^2 \sum_{j=1}^{p} \frac{1}{d_j} = \mathrm{tr}(\mathrm{Var}(\widehat{\boldsymbol{\beta}}^{OLS} \,|\, \mathbb{X}))$$

a

$$\|\mathrm{Bias}(\widehat{\boldsymbol{\beta}}^{Ridge};\, \boldsymbol{\beta} \,|\, \mathbb{X})\|_2^2 = 4\lambda^2\|(\mathbb{X}^\top\mathbb{X} + 2\lambda\mathbb{I}_p)^{-1}\boldsymbol{\beta}\|_2^2 \geq 0 = \|\mathrm{Bias}(\widehat{\boldsymbol{\beta}}^{OLS};\, \boldsymbol{\beta} \,|\, \mathbb{X})\|_2^2.$$

S rostoucím λ variabilita odhadu hřebenovou regresí klesá a euklidovská norma vychýlení roste. Obecně tedy nelze říci, zda nižší podmíněné (vzhledem k pozorovaným vysvětlujícím proměnným $\mathbb{X}$) střední čtvercové chyby dosahuje odhad hřebenovou regresí, či odhad metodou obyčejných nejmenších čtverců. Euklidovská norma vychýlení $\widehat{\beta}_j^{Ridge}(\lambda)$ je rostoucí funkcí $|\beta_j^*|$ a odhady hřebenovou regresí jsou vzhledem k podmíněné střední čtvercové chybě $\mathrm{MSE}(\widehat{\boldsymbol{\beta}},\, \boldsymbol{\beta}^* \,|\, \mathbb{X})$ vhodné zejména v případě malých absolutních hodnot složek vektoru regresních koeficientů $\boldsymbol{\beta}^*$.



(a) $\boldsymbol{\beta}^* = (0, 0, 0, 0, 0)^\top$ (b) $\boldsymbol{\beta}^* = (1, 1, 1, 1, 0)^\top$ (c) $\boldsymbol{\beta}^* = (2, 1, 1, 1, 0)^\top$

Obrázek 1.3: Příklad střední čtvercové chyby odhadu metodou obyčejných nejmenších čtverců a odhadu hřebenovou regresí jako funkce regularizačního parametru λ.

[2] Anglicky: shrinkage

[3] Asymptotická škálovaná střední čtvercová chyba libovolného odhadu může překonat dolní Raovu-Cramérovu mez pouze na množině Lebesguovy míry nula Stoica a Ottersten (1996). V případě řídkých modelů překonáváme Raovu-Cramérovu mez v $\mathbf{0}_p$ (ilustrace na obrázku 1.4).

1.4.1 Vlastnosti žádoucí pro odhady v řídkých modelech

Pro odhady v řídkých lineárních regresních modelech jsou žádoucí následující vlastnosti (první tři uvedené vlastnosti převzaty z Fan a Li (2001), sekce 2).

1. *Skoro-nestrannost* – odhad je málo vychýlený, zejména pro velké absolutní hodnoty β_j^*, $j = 1, \ldots, p$.

2. *Řídkost* – o odhadu vektoru regresních koeficientů řekneme, že je řídký, jestliže *určité množství* regresních koeficientů odhadne jako nulové. Řídkost je vhodná pro automatický výběr relevantních vysvětlujících proměnných a následně snadnější interpretaci odhadnutých koeficientů. Zřejmě je vhodné, aby odhady v řídkém regresním modelu byly řídké.

   Pro řídkost odhadu jsme nalezli více různých definic, Leeb a Pötscher (2008) řídkost odhadu definují[4] jako

   $$\forall \boldsymbol{\beta}^* \in \mathbb{R}^p : \mathbb{P}\left(\mathrm{Supp}(\widehat{\boldsymbol{\beta}}) \subseteq \mathrm{Supp}(\boldsymbol{\beta}^*)\right) \longrightarrow 1, \qquad n \to \infty. \tag{1.21}$$

   Nicméně jedním z hlavních cílů řídkých odhadů je identifikace množiny aktivních složek vektoru regresních koeficientů $\mathrm{Supp}(\boldsymbol{\beta}^*)$, v takovém případě je žádoucí $\mathrm{Supp}(\widehat{\boldsymbol{\beta}}) = \mathrm{Supp}(\boldsymbol{\beta}^*)$ nebo alespoň $\mathrm{Supp}(\widehat{\boldsymbol{\beta}}) \supseteq \mathrm{Supp}(\boldsymbol{\beta}^*)$.

3. *Spojitost* – výsledný odhad je spojitý v pozorovaných datech $\mathbb{X}$ a $\boldsymbol{Y}$. Spojitost odhadu předchází nestabilním odhadům, tedy odhadům, u kterých malá změna v pozorovaných datech vede k velké změně v odhadu.

4. *Výpočetní nenáročnost* – odhad lze vypočítat na základě algoritmu, který nemá vysokou asymptotickou složitost.

5. *Přesnost* – supremum rizika odhadu je blízké supremu rizika *minimaxového odhadu*. Minimaxový odhad parametru $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ je odhad $\widehat{\boldsymbol{\theta}}^M \in \widehat{\boldsymbol{\Theta}} \subseteq \boldsymbol{\Theta}$, který mezi všemi odhady dosahuje nejnižšího rizika v nejhorším možném případě, tedy

   $$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^M) = \inf_{\widehat{\boldsymbol{\theta}} \in \widehat{\boldsymbol{\Theta}}} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}).$$

   V této práci se zabýváme Euklidovskými ztrátovými funkcemi $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2^2$ a $\|\mathbb{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2^2/n$ a riziko je vypočítáno na základě podmíněné střední hodnoty $\mathbb{E}_{\boldsymbol{\beta}^*}[\bullet \,|\, \mathbb{X}]$.

Odhad metodou obyčejných nejmenších čtverců je minimaxovým odhadem v prostoru nestranných lineárních odhadů vzhledem k podmíněné střední čtvercové chybě odhadu. Je spojitý v pozorovaných datech a účelová funkce (1.7) je konvexní[5] v $\boldsymbol{\beta}$, ale není řídký a pro $n < p$ není jednoznačný. Z metod výběr podmnožiny, lasso a hřebenová regrese ani jedna neprodukuje nestranné odhady. Řídké odhady produkují pro vhodnou volbu posloupnosti regularizačních parametrů pouze metody výběr podmnožiny a lasso. Spojité a výpočetně nenáročné odhady produkují pouze lasso a hřebenová regrese. Dle Raskutti a kol. (2011) je pro metodu lasso, za jistých restriktivních předpokladů na matici modelu, *minimaxový řád konvergence*[6] $(p_{\mathcal{S}}/n)\log(p/p_{\mathcal{S}})$ s kladnou pravděpodobností.

[4] Název *sparsity-type condition* v Leeb a Pötscher (2008).

[5] Konvexní problémy jsou běžně řešitelné v polynomiálním čase, výpočetní náročnost metody obyčejných nejmenších čtverců je $\mathcal{O}(n^3)$.

[6] Definici minimaxového řádu konvergence lze nalézt v Wasserman, sekce 3.

Věštecká vlastnost odhadu

Připomeňme značení

$$\mathcal{S} = \mathrm{Supp}(\boldsymbol{\beta}^*) = \left\{ j \in \{1, \ldots, p\} : \beta_j^* \neq 0 \right\} \text{ a } p_{\mathcal{S}} = |\mathcal{S}|.$$

Odhad

$$\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{Oracle} = (\mathbb{X}_S^\top \mathbb{X}_S)^{-1} \mathbb{X}_S^\top \boldsymbol{Y} = \widetilde{\boldsymbol{\beta}}_{\mathcal{S}}^{OLS},$$
$$\widehat{\boldsymbol{\beta}}_{-\mathcal{S}}^{Oracle} = \mathbf{0}_{p-p_{\mathcal{S}}},$$

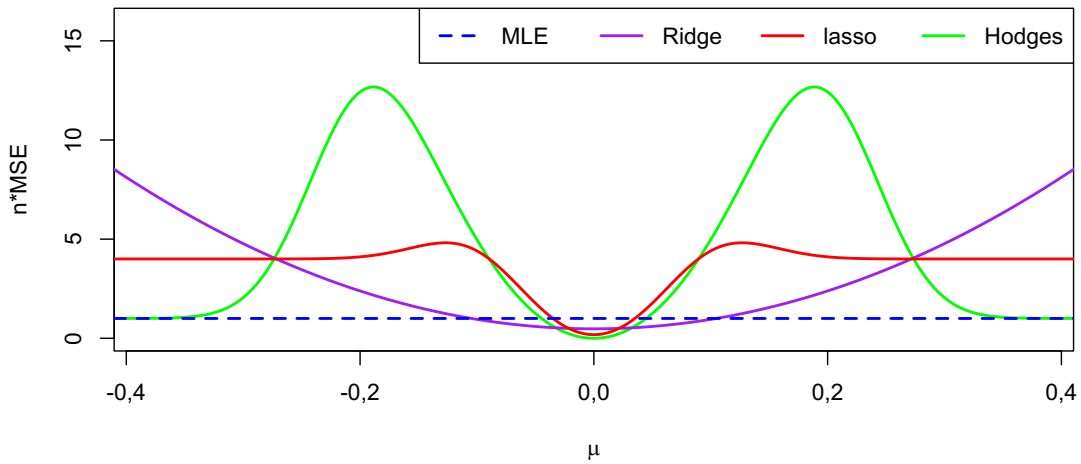
budeme nazývat *věštecký odhad* vektoru regresních koeficientů. Jedná se o odhad metodou obyčejných nejmenších čtverců za předpokladů známé množiny aktivních složek parametru $\boldsymbol{\beta}^*$ a $p_{\mathcal{S}} \leq n-1$. Věštecký odhad má dvě žádoucí vlastnosti, které se nazývají *věštecké vlastnosti odhadu* (Fan a Li (2001), Theorem 2):

1. Odhad $\widehat{\boldsymbol{\beta}}$ asymptoticky přesně identifikuje množinu aktivních složek vektoru regresních koeficientů
$$\mathbb{P}\left(\mathrm{Supp}(\widehat{\boldsymbol{\beta}}) = \mathrm{Supp}(\boldsymbol{\beta}^*)\right) \longrightarrow 1, \qquad n \to \infty. \tag{1.22}$$

2. Pro aktivní složky $\boldsymbol{\beta}^*$ je asymptotické rozdělení odhadu $\widehat{\boldsymbol{\beta}}$ normální, s optimálním asymptotickým rozptylem
$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}^*) \xrightarrow{\mathcal{D}} \mathcal{N}_{p_{\mathcal{S}}}(\mathbf{0}_{p_{\mathcal{S}}}, \sigma^2(\mathbb{E}[\boldsymbol{X}_{\mathcal{S}}\boldsymbol{X}_{\mathcal{S}}^\top])^{-1}), \qquad n \to \infty.$$

Odhad který asymptoticky přesně identifikuje množinu aktivních složek, je zřejmě i řídký a vlastnost (1.22) je více žádoucí než vlastnost (1.21). Řídkost odhadu bývá propagována jako žádoucí. V Leeb a Pötscher (2008), Theorem 2.1 ale dokazují, že libovolný odhad, který splňuje (1.21) má asymptoticky shora neomezenou škálovanou střední čtvercovou chybu

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{E}_{\boldsymbol{\beta}} \left[n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|_2^2 \right] \longrightarrow \infty, \qquad n \to \infty.$$



Obrázek 1.4: Škálovaná střední čtvercová chyba $n\mathbb{E}(\widehat{\mu} - \mu)^2$, $n = 500$, odhadu střední hodnoty $\mu \in \mathbb{R}$ v modelu $\mathcal{N}(\mu, 1)$. Použity maximálně věrohodný odhad (MLE) $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, analogie hřebenové regrese (Ridge) $\overline{X}_n/\lambda$ s volbou regularizačního parametru $\lambda = 10/\sqrt{n}$, analogie metody lasso (lasso) $S_{(2/n^{1/2})}(\overline{X}_n)$, kde S je měkká prahovací funkce a Hodgesův odhad (Hodges) $H_{(1/n^{1/4})}(\overline{X}_n)$, kde H je tvrdá prahovací funkce, čili speciální analogie výběru podmnožiny. Kvůli „různému chování odhadů vzhledem k volbě regularizačního parametru“ obrázek výše neporovnává kvality jednotlivých odhadů, ale pouze tvary škálované střední čtvercové chyby.

2. Metoda lasso

Tato kapitola je věnována metodě lasso z Tibshirani (1996). V sekci 2.1 je odvozena existence, tvar a podmínky pro jednoznačnost odhadu metodou lasso. V sekci 2.2 je uveden algoritmus pro výpočet tohoto odhadu. V sekcích 2.1 a 2.2 čerpáme z Tibshirani (2013). Sekce 2.3 je věnována geometrickému a Bayesovskému pohledu na metodu lasso. V sekci 2.4 jsou uvedeny asymptotické vlastnosti odhadu metodou lasso a čerpáme zde z Fu a Knight (2000) a Zou (2006).

2.1 Existence, tvar a jednoznačnost

V metodě lasso je v úloze nejmenších čtverců (1.7) aplikována penalizační funkce $\lambda\|\boldsymbol{\beta}\|_1$, $\lambda \geq 0$ na vektor $\boldsymbol{\beta}$. Brzy uvidíme, proč penalizační funkce $\lambda\|\boldsymbol{\beta}\|_1$ vede k produkci řídkých odhadů.

Definice 2.1 (Odhad metodou lasso)**.** *Nechť n, $p \in \mathbb{N}$, $\lambda \in \mathbb{R}$, $\lambda \geq 0$, $\boldsymbol{Y}$ je n-rozměrný vektor odezev a $\mathbb{X}$ je matice modelu typu $n \times p$. Libovolné řešení úlohy*

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \qquad \lambda \geq 0. \tag{2.1}$$

budeme značit $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ a nazývat odhad metodou lasso[1]*.*

Existence řešení úlohy (2.1) plyne ze standardní teorie konvexního programování – pro $\lambda > 0$ je účelová funkce úlohy (2.1) konvexní (lemma A.1.1 a A.1.2) a bez směru neomezeného poklesu, nabývá tedy svého minima na $\mathbb{R}^p$. Níže uvádíme lemma o řešení úlohy (2.1) z Tibshirani (2013). Ekvivalentní tvrzení pro vázanou formu lze rovněž nalézt v Osborne a kol. (2000a), Theorem 1.

Lemma 2.2 (Tibshirani (2013), Lemma 1)**.** *Pro libovolné $\boldsymbol{Y} \in \mathbb{R}^p$, $\mathbb{X} \in \mathbb{R}^{n\times p}$ a $\lambda \geq 0$ má lasso následující vlastnosti:*

1. *Úloha (2.1) má buď právě jedno řešení, nebo nespočetně mnoho řešení.*
2. *Hodnota $\mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ je pro všechna řešení $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ stejná.*
3. *Je-li $\lambda > 0$, pak mají všechna řešení (2.1) shodnou ℓ^1-normu $\|\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\|_1$.*

Účelová funkce úlohy (2.1) je konvexní, ale není parciálně diferencovatelná v bodě 0. Tvar řešení nalezneme pomocí lemmatu 1.3 (subgradientní podmínka optimality), lemmatu 1.4 (o subdiferenciálu diferencovatelné konvenxní funkce) a lemmatu 2.2. Subdiferenciál účelové funkce úlohy (2.1) v bodě $\boldsymbol{\beta} \in \mathbb{R}^p$ je

$$\partial f(\boldsymbol{\beta};\, \mathbb{X},\, \boldsymbol{Y},\, \lambda) = \{-\mathbb{X}^\top (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) + \lambda\boldsymbol{s} : \boldsymbol{s} \in \partial\|\boldsymbol{\beta}\|_1\},$$

kde složky vektoru $\boldsymbol{s} = (s_1, \ldots, s_p)^\top \in \partial\|\boldsymbol{\beta}\|_1$ splňují

$$s_j \in \begin{cases} \{1\}, & \beta_j > 0, \\ [-1,\, 1], & \beta_j = 0, \\ \{-1\}, & \beta_j < 0, \end{cases} \qquad j = 1, \ldots, p. \tag{2.2}$$

[1] Lasso je zkratkou z anglického *least absolute shrinkage and selection operator*.

Subgradientní podmínka optimality pro účelovou funkci úlohy (2.1) je tedy

$$\mathbb{X}^\top (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) = \lambda \boldsymbol{s}(\lambda), \qquad \boldsymbol{s}(\lambda) \in \partial \|\boldsymbol{\beta}\|_1. \tag{2.3}$$

Podmínka (2.3) je splněna pro odhad metodou lasso $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$, neboť se jedná o řešení úlohy (2.1), které dle lemmatu 2.2 existuje. Na $\boldsymbol{s}(\lambda)$ které splňuje subgradientní podmínku optimality (2.3) lze tedy nahlížet jako na funkci $\mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$. Hodnota $\mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ je dle lemmatu 2.2 určena jednoznačně, tedy i $\boldsymbol{s}(\lambda)$ je určeno jednoznačně a z (2.2), resp. (2.3) pro aktivní složky odhadu plyne

$$s_j(\lambda) = \operatorname{sign}(\widehat{\beta}_j^{Lasso}(\lambda)), \qquad j \in \operatorname{supp}(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)),$$

resp.

$$s_j(\lambda) = \operatorname{sign}(\boldsymbol{X}_j^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))), \qquad j \in \operatorname{supp}(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)). \tag{2.4}$$

Z podmínek (2.2) a (2.3) lze odvodit i tvar množiny aktivních složek odhadu metodou lasso $\mathcal{A}(\lambda) = \operatorname{supp}(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))$, která je rovněž určena jednoznačně, neboť je rovněž funkcí $\mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$, respektive $\boldsymbol{s}(\lambda)$

$$\mathcal{A}(\lambda) = \{j : |\boldsymbol{X}_j^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))| = \lambda\} = \{j : |s_j(\lambda)| = 1\}. \tag{2.5}$$

Množina $\mathcal{A}(\lambda)$ bývá též nazývána *ekvikorelační množinou*[2], pravděpodobně neboť za předpokladu studentizovaných vysvětlujících proměnných platí pro $j \in \mathcal{A}(\lambda)$:

1. $|\boldsymbol{X}_j^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))|$ je proporcionální absolutní hodnotě kosinu úhlu mezi náhodným vektorem $\boldsymbol{X}_j$ a vektorem reziduí odhadu metodou lasso

   $$|\widehat{\rho}_j(\lambda)| := \frac{|\boldsymbol{X}_j^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))|}{\|\boldsymbol{X}_j\|_2 \cdot \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\|_2} = \frac{|\boldsymbol{X}_j^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))|}{\sqrt{n-1}\|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\|_2}.$$

   Pro vysvětlující proměnné s nulovým výběrovým průměrem zřejmě platí $\boldsymbol{X}_j = (\boldsymbol{X}_j - \overline{X}_{\bullet,j} \cdot \boldsymbol{1}_n)$, kde $\overline{X}_{\bullet,j} = \frac{1}{n}\sum_{i=1}^n X_{i,j}$ je výběrový průměr náhodného výběru z j-té vysvětlující proměnné X_j. Hodnotu $\widehat{\rho}_j(\lambda)$ lze tedy považovat za analogii výběrového Pearsonova korelačního koeficientu mezi náhodnými veličinami X_j a Y, kde je odhad střední hodnoty $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ nahrazen odhadem podmíněné střední hodnoty $\mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$.

2. Hodnota $|\widehat{\rho}_j(\lambda)|$ je pro všechna $j \in \mathcal{A}(\lambda)$ stejná (plyne z (2.5)) a za podmínky (2.3) maximální možná.

3. Vhodnější termín „stejnoúhlý[3]“ je již použit v Efron a kol. (2004) s mírně odlišným významem.

Nyní odvodíme tvar řešení odhadu metodou lasso. Z definice množiny aktivních složek odhadu $\mathcal{A}(\lambda)$ zřejmě platí $\widehat{\boldsymbol{\beta}}^{Lasso}_{-\mathcal{A}(\lambda)} = \boldsymbol{0}_{-\mathcal{A}(\lambda)}$ a pro aktivní složky odhadu $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ lze řešení podmínky (2.3) vyjádřit jako

$$\boldsymbol{X}_j^\top (\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda)}\widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda)}(\lambda)) = \lambda s_j(\lambda), \qquad j \in \mathcal{A}(\lambda), \tag{2.6}$$

čili

$$\mathbb{X}_{\mathcal{A}(\lambda)}^\top (\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda)}\widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda)}(\lambda)) = \lambda \boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda). \tag{2.7}$$

[2] Anglicky: equicorrelation set

[3] Anglicky: equiangular

Z (2.7) plyne $\lambda \boldsymbol{s}_{\mathcal{A}} \in \mathbb{X}_{\mathcal{A}}^{\top}$ a tedy $\lambda \boldsymbol{s}_{\mathcal{A}} = \mathbb{X}_{\mathcal{A}}^{\top}(\mathbb{X}_{\mathcal{A}}^{\top})^{+}\lambda \boldsymbol{s}_{\mathcal{A}}$. K výpočtu řešení soustavy (2.7) použijeme analogický postup, jako byl použit v sekci 1.2. Jelikož

$$(\mathbb{X}_{\mathcal{A}}^{\top}\mathbb{X}_{\mathcal{A}})(\mathbb{X}_{\mathcal{A}}^{\top}\mathbb{X}_{\mathcal{A}})^{+}(\mathbb{X}_{\mathcal{A}}^{\top}\boldsymbol{Y} - \lambda \boldsymbol{s}_{\mathcal{A}}) = \mathbb{X}_{\mathcal{A}}^{\top}(\boldsymbol{Y} - \lambda(\mathbb{X}_{\mathcal{A}}^{\top})^{+}\boldsymbol{s}_{\mathcal{A}}) = \mathbb{X}_{\mathcal{A}}^{\top}\boldsymbol{Y} - \lambda \boldsymbol{s}_{\mathcal{A}}$$

všechna řešení soustavy $(\mathbb{X}_{\mathcal{A}}^{\top}\mathbb{X}_{\mathcal{A}})\boldsymbol{\beta} = \mathbb{X}_{\mathcal{A}}^{\top}\boldsymbol{Y} - \lambda \boldsymbol{s}_{\mathcal{A}}$ lze charakterizovat vlastnostmi

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathcal{A}(\lambda)}^{Lasso}(\lambda) \in &\{(\mathbb{X}_{\mathcal{A}(\lambda)}^{+}\boldsymbol{Y} - \lambda(\mathbb{X}_{\mathcal{A}(\lambda)}^{\top}\mathbb{X}_{\mathcal{A}(\lambda)})^{+}\boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda)) + \boldsymbol{\gamma},\, \boldsymbol{\gamma} \in \operatorname{Ker}(\mathbb{X}_{\mathcal{A}(\lambda)})\} \\ &\& \operatorname{sign}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}(\lambda)}^{Lasso}(\lambda)) = \boldsymbol{s}_{\mathcal{A}(\lambda)}.\end{aligned} \tag{2.8}$$

Člen $\mathbb{X}_{\mathcal{A}(\lambda)}^{+}\boldsymbol{Y} = \widetilde{\boldsymbol{\beta}}_{\mathcal{A}(\lambda)}^{OLS}$ v (2.8) odpovídá odhadu vektoru regresních koeficientů metodou obyčejných nejmenších čtverců s nejmenší euklidovskou normou v modelu $\mathcal{A}(\lambda)$. Člen $\lambda(\mathbb{X}_{\mathcal{A}(\lambda)}^{\top}\mathbb{X}_{\mathcal{A}(\lambda)})^{+}\boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda)$ je „zmenšovací člen[4]“. Za předpokladu plné sloupcové hodnosti matice $\mathbb{X}_{\mathcal{A}(\lambda)}$ je řešení soustavy (2.7) jednoznačné a tvaru

$$\widehat{\boldsymbol{\beta}}_{\mathcal{A}(\lambda)}^{Lasso} = (\mathbb{X}_{\mathcal{A}(\lambda)}^{\top}\mathbb{X}_{\mathcal{A}(\lambda)})^{-1}(\mathbb{X}_{\mathcal{A}(\lambda)}^{\top}\boldsymbol{Y} - \lambda \boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda)). \tag{2.9}$$

Postačující podmínkou k plné sloupcové hodnosti matice $\mathbb{X}_{\mathcal{A}(\lambda)}$ je *obecná poloha* sloupců matice modelu $\mathbb{X}$ v afinním prostoru dimenze alespoň $|\mathcal{A}(\lambda)|$. Uvádíme modifikaci definice obecné polohy sloupců matice z Tibshirani (2013).

Definice 2.3. *Nechť d, n, $p \in \mathbb{N}$, $d \leq n$, $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p) \in \mathbb{R}^{n\times p}$ je matice a $\mathcal{J}_d \subseteq \mathbb{R}^n$ je afinní prostor dimenze d. O matici $\mathbb{X}$ řekneme, že má sloupce v obecné poloze v prostoru $\mathcal{J}_d$, jestliže $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p \in \mathcal{J}_d$ a pro libovolné $k \in \mathbb{N}$, $k < \min\{d, p\}$ žádný k-dimenzionální afinní podprostor $\mathcal{J} \subset \mathcal{J}_d$ neobsahuje více než $k+1$ prvků množiny $\{\pm\boldsymbol{X}_1, \ldots, \pm\boldsymbol{X}_p\}$, až na protilehlé páry.*

Alternativně pro libovolnou volbu znamének $\sigma_1, \ldots, \sigma_{k+1} \in \{-1, 1\}$ a libovolnou volbu indexů $\{i_1, \ldots, i_{k+1}\} \subseteq \{1, \ldots, p\}$ platí

$$\operatorname{Aff}(\sigma_1\boldsymbol{X}_{i_1}, \ldots, \sigma_{k+1}\boldsymbol{X}_{i_{k+1}}) \bigcap \{\pm\boldsymbol{X}_i : i \neq i_1, \ldots, i_{k+1}\} = \emptyset, \tag{2.10}$$

kde $\operatorname{Aff}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k+1}) = \{\sum_{j=1}^{k+1}\alpha_j\boldsymbol{X}_j : \alpha_j \in \mathbb{R},\, j = 1, \ldots, k+1,\, \sum_{j=1}^{k+1}\alpha_j = 1\}$ je afinní obal posloupnosti vektorů $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k+1}$.

Poznámka. Nechť n, $p \in \mathbb{N}$, $n < p$ a $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p) = \mathbb{X} \in \mathbb{R}^{n\times p}$ je matice s centrovanými sloupci, tedy $\mathbb{X} = (\mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\top})\mathbb{X}$. Sloupce matice $\mathbb{X}$ nemohou být v obecné poloze v prostoru $\mathbb{R}^n$, neboť $\operatorname{Im}(\mathbb{X}) \subseteq \operatorname{Im}(\mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\top})$. Tedy existuje afinní prostor $\mathcal{J} = \operatorname{Im}(\mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\top}) \subseteq \mathbb{R}^n$ dimenze $k = n - 1 < \min\{n, p\}$, který obsahuje $p > k+1$ prvků množiny $\{\pm\boldsymbol{X}_1, \ldots, \pm\boldsymbol{X}_p\}$. Sloupce matice $\mathbb{X}$ stále mohou být v obecné poloze v afinním prostoru $\operatorname{Im}(\mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\top})$.

Nechť $\mathcal{J}_d \subseteq \mathbb{R}^n$ je afinní prostor dimenze $d \leq n$ a $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p \in \mathcal{J}_d$. Drobným rozšířením Tibshirani (2013), diskuze nad lemma 4, postačující podmínkou pro skoro jistě obecnou polohu sloupců matice $\mathbb{X}$ v afinním prostoru $\mathcal{J}_d$ je $\mathbb{P}([\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p \in \mathcal{J}_d] = 1)$ a absolutní spojitost sdruženého rozdělení sloupců matice $\mathbb{X}$ vzhledem k Lebesgueově míře v $\mathcal{J}_d^p$. V takovém případě pro $k < \min\{d, p\}$

$$\mathbb{P}(\sigma_{k+2}\boldsymbol{X}_{k+2} \in \operatorname{Aff}(\sigma_1\boldsymbol{X}_1, \ldots, \sigma_{k+1}\boldsymbol{X}_{k+1}) \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k+1}) = 0, \tag{2.11}$$

neboť pro pevně dané $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k+1} \in \mathcal{J}_d$ má množina $\operatorname{Aff}(\sigma_1\boldsymbol{X}_1, \ldots, \sigma_{k+1}\boldsymbol{X}_{k+1})$ v prostoru $\mathcal{J}_d$ Lebesgueovu míru 0. Vyintegrováním$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k+1}$ v (2.11) poté získáváme i $\mathbb{P}(\sigma_{k+2}\boldsymbol{X}_{k+2} \in \operatorname{Aff}(\sigma_1\boldsymbol{X}_1, \ldots, \sigma_{k+1}\boldsymbol{X}_{k+1})) = 0$. Sjednocením přes všechny $(k+2)$-prvkové podmnožiny posloupnosti $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, všechny variace $k+2$ znamének $(\sigma_1, \ldots, \sigma_{k+2})$, $\sigma_1, \ldots, \sigma_{k+2} \in \{-1, 1\}$ a všechna $k < \min\{d, p\}$ získáváme, že jev (2.10) nastane skoro jistě.

[4] Anglicky: *shrinkage term*

2.2 Výpočetní algoritmus

Odhad $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$, $\lambda \in [0, \infty)$ lze vypočítat iterativním řešením subgradientní podmínky optimality (2.3) pro λ klesající od ∞ do 0 (hodnota ∞ je zde míněna jako hodnota, která pro vstupní data $\boldsymbol{Y}$, $\mathbb{X}$, splňuje $\widehat{\boldsymbol{\beta}}^{Lasso}(\infty) = \mathbf{0}_p$). Algoritmus, který iterativně řeší (2.3), se nazývá *modifikace algoritmu LARS pro lasso*. Algoritmus LARS[5] i jeho modifikace pro lasso byly představeny v Efron a kol. (2004). Modifikace algoritmu LARS pro lasso uvedena v Efron a kol. (2004) předpokládá $\text{rank}(\mathbb{X}_{\mathcal{A}}) = |\mathcal{A}|$, což nemusí vždy platit. Uvádíme proto algoritmus „*The LARS algorithm for the lasso path*“ z Tibshirani (2013), který funguje i v případě $\text{rank}(\mathbb{X}_{\mathcal{A}}) < |\mathcal{A}|$. Níže je schématicky zapsaný algoritmus a jeho podrobnější vysvětlení. Kompletní algoritmus je uveden v apendixu (Algoritmus 1).

0. Inicializace proměnných:
   - Počítadlo iterací $k := 0$.
   - Počáteční uzel $\lambda_0 := \infty$.
   - Počáteční množina aktivních složek odhadu $\mathcal{A}(\lambda_0) := \emptyset$.
   - Počáteční vektor aktivních znamének odhadu $\boldsymbol{s}_{\mathcal{A}(\lambda_0)}(\lambda_0) := \emptyset$.
1. Dokud je $\lambda_k > 0$:
   - Výpočet řešení $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_k)$ podle vzorce (2.12).
   - Výpočet uzlu λ_{k+1}^{join}, prvního vstupu nové složky $j \in \{1, \ldots, p\} \setminus \mathcal{A}(\lambda_k)$ do množiny aktivních složek odhadu využitím vzorců (2.15) a (2.16).
   - Výpočet uzlu λ_{k+1}^{leave}, prvního výstupu složky $j \in \mathcal{A}(\lambda_k)$ z množiny aktivních složek odhadu využitím vzorců (2.17) a (2.18).
   - Další uzel $\lambda_{k+1} := \max\{\lambda_{k+1}^{join}, \lambda_{k+1}^{leave}\}$.
   - Aktualizace množiny aktivních složek $\mathcal{A}(\lambda_{k+1})$ a aktivních znamének $\boldsymbol{s}_{\mathcal{A}(\lambda_{k+1})}(\lambda_{k+1})$ pomocí (2.19) v případě $\lambda_{k+1} = \lambda_{k+1}^{join}$, respektive pomocí (2.20) v případě $\lambda_{k+1} = \lambda_{k+1}^{leave}$.
   - Aktualizace počítadla iterací $k := k + 1$.

Výsledkem algoritmu je spojitá, po částech afinní funkce, která parametru λ přiřadí hodnotu odhadu $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$. Počet iterací algoritmu je shora omezen konstantou $(3^p + 1)/2$, neboť:

- Každá dvojice množiny aktivních složek odhadu a aktivních znamének odhadu $(\mathcal{A}, \boldsymbol{s}_{\mathcal{A}})$ se vyskytne nejvýše jednou (Osborne a kol. (2000b), Property 1). Toto dává horní mez 3^p pro počet iterací algoritmu.
- Vyskytne-li se během algoritmu dvojice $(\mathcal{A}, \boldsymbol{s}_{\mathcal{A}})$, nemůže se již vyskytnout dvojice $(\mathcal{A}, -\boldsymbol{s}_{\mathcal{A}})$ (Mairal a Yu (2012), Proposition 1). Toto snižuje horní mez pro počet iterací algoritmu na $(3^p + 1)/2$.
- Navíc pro nestandardizované vysvětlující proměnné má Algoritmus 1 v nejhorším případě právě $(3^p + 1)/2$ iterací (Mairal a Yu (2012), Theorem 1).

[5]LARS je zkratkou z anglického *„least angle regression“*, S značí *„stagewise“* a *„lasso“*.

Výpočet řešení: Odhad metodou lasso $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ je v případě $|\mathcal{A}(\lambda)| \leq \text{rank}(\mathbb{X})$ dán vztahem (2.8). Je-li $|\mathcal{A}(\lambda)| > \text{rank}(\mathbb{X})$ není $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ určeno jednoznačně. V Tibshirani (2013) z množiny řešení vybírají řešení s nejmenší euklidovskou normou. Odhad $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ je tedy dán vztahem

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda)}(\lambda) &:= \mathbb{X}^{+}_{\mathcal{A}(\lambda)}\boldsymbol{Y} - \lambda(\mathbb{X}^{\top}_{\mathcal{A}(\lambda)}\mathbb{X}_{\mathcal{A}(\lambda)})^{+}\boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda), \\ \widehat{\boldsymbol{\beta}}^{Lasso}_{-\mathcal{A}(\lambda)}(\lambda) &:= \mathbf{0}_{-\mathcal{A}(\lambda)}. \end{aligned} \tag{2.12}$$

Výpočet uzlů a znamének: Během zmenšování λ je třeba dohlédnout na dodržení subgradientní podmínky optimality (2.3):

$$|\boldsymbol{X}_j^{\top}(\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda_k)}\widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_k)}(\lambda))| < \lambda, \qquad \lambda < \lambda_k,\, j \in \mathcal{A}^C(\lambda_k), \tag{2.13}$$

a

$$\text{sign}(\widehat{\boldsymbol{\beta}}^{Lasso}_j(\lambda)) = s_j(\lambda_k), \qquad \lambda < \lambda_k,\ j \in \mathcal{A}(\lambda_k), \tag{2.14}$$

kde $\mathcal{A}^C(\lambda_k) = \{1, \ldots, p\} \setminus \mathcal{A}(\lambda_k)$. V případě porušení (2.13) nebo (2.14) je nutné odhad (2.12) aktualizovat tak, aby znaménka aktualizovaného odhadu podmínku (2.3) splňovala.

Uvažujme k-tou iteraci Algoritmu 1. Pro j-tou složku, $j \in \mathcal{A}^C(\lambda_k)$, dojde ze spojitosti $\lambda \mapsto \widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_k)}(\lambda)$ k prvnímu porušení podmínky (2.13) v případě

$$\boldsymbol{X}_j^{\top}(\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda_k)}\widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_k)}(\lambda)) = \pm\lambda, \qquad \lambda < \lambda_k,\, j \in \mathcal{A}^C(\lambda_k),$$

což lze rozepsat jako

$$\boldsymbol{X}_j^{\top}(\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda_k)}\widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(\lambda_k)} + \lambda(\mathbb{X}^{\top}_{\mathcal{A}(\lambda_k)})^{+}\boldsymbol{s}(\lambda_k)) = \pm\lambda, \qquad \lambda < \lambda_k,\, j \in \mathcal{A}^C(\lambda_k).$$

Vyjádřením λ získáváme uzel, ve kterém by j-tá složka, $j \in \mathcal{A}^C(\lambda_k)$, vstoupila do množiny aktivních složek odhadu

$$t_j^{join} \in \left\{ \frac{\boldsymbol{X}_j^{\top}(\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda_k)}\widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(\lambda_k)})}{\pm 1 - \boldsymbol{X}_j^{\top}(\mathbb{X}^{\top}_{\mathcal{A}(\lambda_k)})^{+}\boldsymbol{s}(\lambda_k)} \right\} \bigcap [0,\, \lambda_k), \qquad j \in \mathcal{A}^C(\lambda_k). \tag{2.15}$$

První porušení podmínky (2.13) nastane v

$$\lambda^{join}_{k+1} := \max_{j \in \mathcal{A}^C(\lambda_k)} t_j^{join}. \tag{2.16}$$

Příslušný index a znaménko jsou

$$j^{join}_{k+1} := \operatorname*{argmax}_{j \in \mathcal{A}^C(\lambda_k)} t_j^{join}, \qquad s^{join}_{k+1} := \text{sign}(\boldsymbol{X}^{\top}_{j^{join}_{k+1}}(\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda_k)}\widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_k)}(\lambda^{join}_{k+1}))).$$

Pro j-tou složku, $j \in \mathcal{A}(\lambda_k)$, dojde k prvnímu porušení podmínky (2.14) v případě

$$(\widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(\lambda_k)} - \lambda(\mathbb{X}^{\top}_{\mathcal{A}(\lambda_k)}\mathbb{X}_{\mathcal{A}(\lambda_k)})^{+}\boldsymbol{s}(\lambda_k))_j = 0, \qquad \lambda < \lambda_k,\, j \in \mathcal{A}(\lambda_k).$$

Vyjádřením λ získáváme uzel, ve kterém by j-tá složka, $j \in \mathcal{A}(\lambda_k)$, vystoupila z množiny aktivních složek odhadu

$$t_j^{leave} = \frac{(\widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(\lambda_k)})_j}{([\mathbb{X}^{\top}_{\mathcal{A}(\lambda_k)}\mathbb{X}_{\mathcal{A}(\lambda_k)}]^{+}\boldsymbol{s}(\lambda_k))_j} \cdot \mathbb{I}\left[\frac{(\widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(\lambda_k)})_j}{([\mathbb{X}^{\top}_{\mathcal{A}(\lambda_k)}\mathbb{X}_{\mathcal{A}(\lambda_k)}]^{+}\boldsymbol{s}(\lambda_k))_j} < \lambda_k \right], \quad j \in \mathcal{A}(\lambda_k). \tag{2.17}$$

K prvnímu porušení podmínky (2.14) dojde u uzlu

$$\lambda_{k+1}^{leave} := \max_{j \in \mathcal{A}(\lambda_k)} t_j^{leave}. \tag{2.18}$$

Příslušný index a znaménko jsou

$$j_{k+1}^{leave} := \operatorname*{argmax}_{j \in \mathcal{A}(\lambda_k)} t_j^{leave}, \qquad s_{k+1}^{leave} := s_{j_{k+1}^{leave}}(\lambda_k).$$

Tvar odhadu je třeba aktualizovat v uzlu

$$\lambda_{k+1} := \max\{\lambda_k^{join}, \lambda_k^{leave}\}.$$

V případě $\lambda_{k+1} := \lambda_k^{join}$ je

$$\begin{aligned} \mathcal{A}(\lambda_{k+1}) &:= \mathcal{A}(\lambda_k) \cup \{j_{k+1}^{join}\}, \\ \boldsymbol{s}(\lambda_{k+1}) &:= \operatorname{sign}(\mathbb{X}_{\mathcal{A}(\lambda_{k+1})}^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_{\mathcal{A}(\lambda_k)}^{Lasso}(\lambda_{k+1}))). \end{aligned} \tag{2.19}$$

V případě $\lambda_{k+1} = \lambda_k^{leave}$ je

$$\begin{aligned} \mathcal{A}(\lambda_{k+1}) &:= \mathcal{A}(\lambda_k) \setminus \{j_{k+1}^{leave}\}, \\ \boldsymbol{s}(\lambda_{k+1}) &:= \operatorname{sign}(\mathbb{X}_{\mathcal{A}(\lambda_{k+1})}^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_{\mathcal{A}(\lambda_k)}^{Lasso}(\lambda_{k+1}))). \end{aligned} \tag{2.20}$$

Z tvaru řešení (2.12) je zřejmé, že funkce $\lambda \mapsto \widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ je po částech afinní. Důkaz spojitosti odhadu $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ je uveden v (Tibshirani (2013), Lemma 17).

Poznámka. Poukažme na iterativní povahu řešení. Nechť $|\mathcal{A}(\lambda)| \leq \operatorname{rank}(\mathbb{X})$ a porovnejme rovnici (2.4) pro znaménka odhadu metodou lasso se způsobem výpočtu znamének v algoritmu. Vyjádřením $\boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda)$ z rovnice (2.4) získáváme

$$\begin{aligned} \boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda) &= \operatorname{sign}(\mathbb{X}_{\mathcal{A}(\lambda)}^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))) \\ &= \operatorname{sign}[\mathbb{X}_{\mathcal{A}(\lambda)}^\top (\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda)}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}(\lambda)}^{OLS}) + \lambda \mathbb{X}_{\mathcal{A}(\lambda)}^\top \mathbb{X}_{\mathcal{A}(\lambda)} (\mathbb{X}_{\mathcal{A}(\lambda)}^\top \mathbb{X}_{\mathcal{A}(\lambda)})^{-1} \boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda))] \\ &= \operatorname{sign}[\boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda))] \\ &= \boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda), \end{aligned}$$

což nepůsobí příliš užitečně. Algoritmus při výpočtu znamének $\boldsymbol{s}_{\mathcal{A}(\lambda_{k+1})}(\lambda_{k+1})$ v (2.19) a (2.20) využívá odhad $\widehat{\boldsymbol{\beta}}_{\mathcal{A}(\lambda_k)}^{Lasso}(\lambda_{k+1})$ z předchozí iterace, tedy

$$\boldsymbol{s}_{\mathcal{A}(\lambda_{k+1})}(\lambda_{k+1}) = \operatorname{sign}(\mathbb{X}_{\mathcal{A}(\lambda_{k+1})}^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_{\mathcal{A}(\lambda_k)}^{Lasso}(\lambda_{k+1}))).$$

Toto je možné díky spojitosti odhadu $\widehat{\boldsymbol{\beta}}^{Lasso}$ v λ.

2.3 Geometrický a Bayesovský pohled

Pro lepší pochopení různých konceptů je vhodné na tyto koncepty umět nazírat z více různých úhlů. Za cenu vyšších nároků na čas získáváme fluidnější myšlení v daném okruhu. Zde uvádíme geometrický a Bayesovský pohled na metodu lasso. Začlenění těchto perspektiv není novým konceptem, ale nápady prezentované v této části byly převážně vlastní[6].

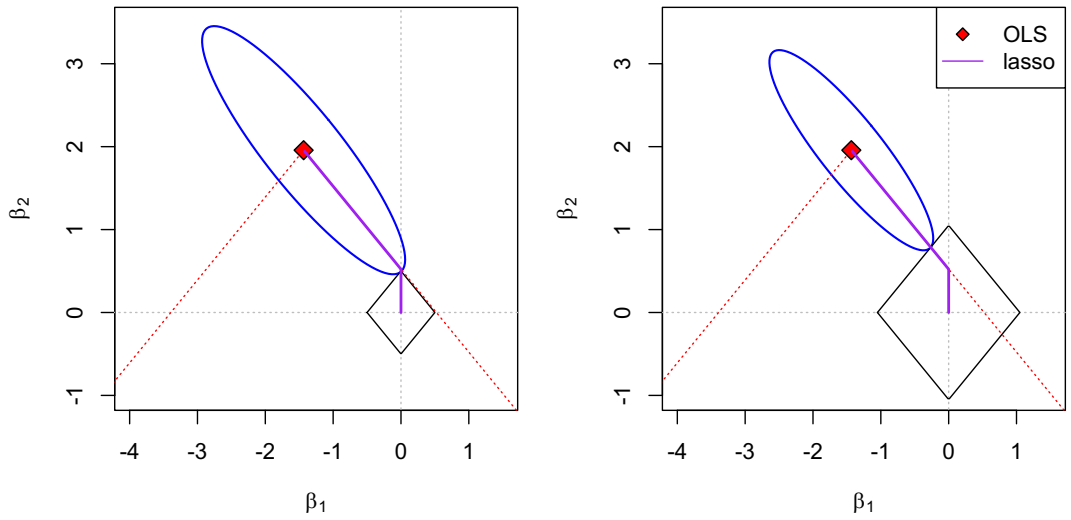
[6] V pozdější části prací jsme zjistili, že níže prezentované nápady byly již dříve publikovány.

2.3.1 Geometrický pohled

Uvažujme lasso (2.1) ve vázané formě

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2, \qquad \text{s.t. } \|\boldsymbol{\beta}\|_1 \leq t.$$

Pro $t \geq 0$ dostatečně malé je řešení ve shodě s intuicí, že se hyperelipsoid dotkne ℓ_1-koule na jednom z vrcholů (respektive hran, stěn, atd.). Na prvním grafu na Obrázku (2.1) je první složka odhadu nulová. Na druhém grafu je již t dostatečně velké (odpovídá dostatečně malému λ v (2.1)) a obě složky odhadu jsou nenulové.



Obrázek 2.1: Lasso v $\mathbb{R}^2$ ve vázané formě pro dvě různé hodnoty parametru $t \geq 0$. Černý kosočtverec vyznačuje hranici množiny přípustných řešení $\|\boldsymbol{\beta}\|_1 \leq t$. Červený bod značí hodnotu odhadu metodou obyčejných nejmenších čtverců. Modré elipsy odpovídají různým vrstevnicím funkce $\boldsymbol{\beta} \mapsto \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{OLS})^\top(\mathbb{X}^\top\mathbb{X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{OLS})$. Červené tečkované polopřímky znázorňují zmenšovací členy pro $|\mathcal{A}| = 2$, tedy $\lambda(\mathbb{X}^\top\mathbb{X})^{-1}(1,\,1)^\top$, respektive $\lambda(\mathbb{X}^\top\mathbb{X})^{-1}(-1,\,1)^\top$ a jsou na sebe kolmé. Fialová křivka odpovídá cestě odhadu $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$. Hodnoty řešení metodou lasso pro konkrétní volbu t odpovídají bodům dotyku černého kosočtverce a modré elipsy.

Příklad (Patologický případ lasso v druhé dimenzi). Uvažujme data sestrojená na základě (Mairal a Yu (2012), Proposition 2 – Adversarial Strategy)

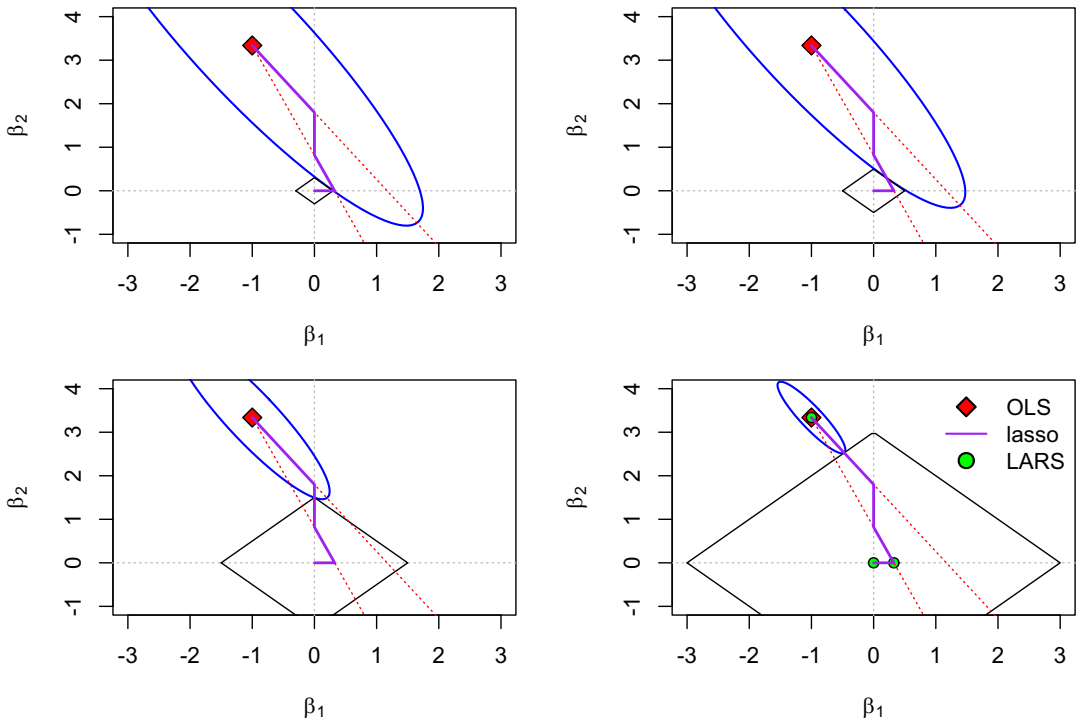
$$\boldsymbol{Y} = \begin{pmatrix} 1{,}00 \\ 0{,}95 \end{pmatrix}, \mathbb{X} = \begin{pmatrix} 1 & 0{,}59 \\ 0 & 0{,}28 \end{pmatrix}. \tag{2.21}$$

Dle (Mairal a Yu (2012), Theorem 1) bude mít cesta odhadu metodou lasso právě $(3^p + 1)/2$ uzlů. Upozorněme ovšem, že data (2.21) nejsou centrovaná a sloupce matice $\mathbb{X}$ nejsou studentizované. Metoda lasso funguje i pro necentrovaná data, ale je vhodné do modelu zahrnout absolutní člen. Zahrnutí absolutního členu do modelu vede k porušení speciální struktury matice $\mathbb{X}$ sestrojené na základě (Mairal a Yu (2012), Proposition 2) a cesta odhadu metodou lasso již nemusí mít maximální možný počet uzlů.

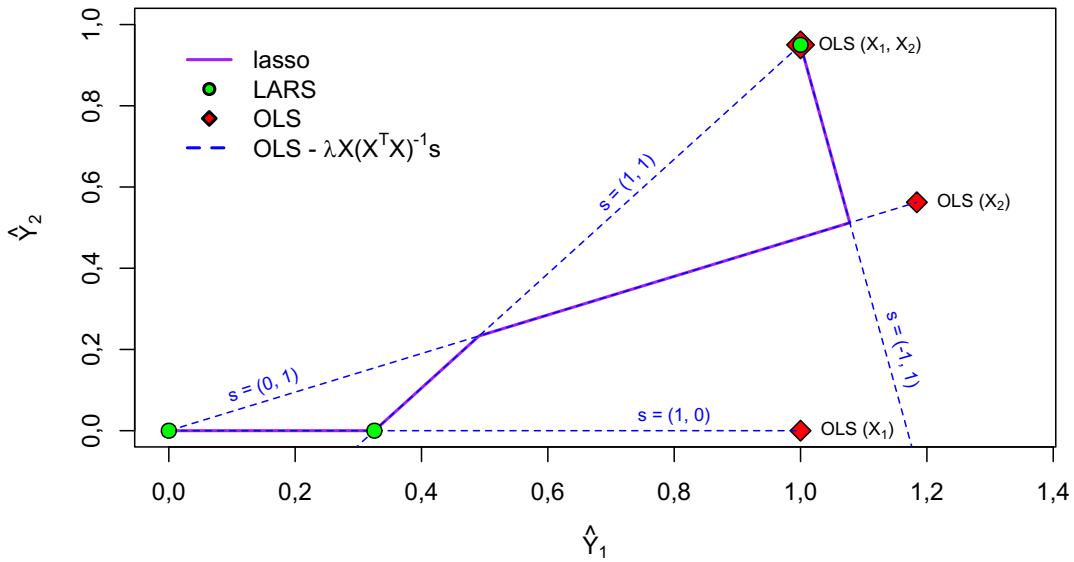
Pro necentrované vysvětlující proměnné není pro aktivní složky odhadu hodnota

$$\lambda = |\boldsymbol{X}_j^\top(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))|, \qquad j \in \operatorname{supp}(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))$$

proporcionální analogii empirického Pearsonova korelačního koeficientu (zmíněné v sekci 2.1) náhodných veličin X_j a Y.



Obrázek 2.2: Lasso v $\mathbb{R}^2$ s maximálním možným počtem uzlů pro čtyři různé, rostoucí hodnoty parametru $t \geq 0$. Černý kosočtverec vyznačuje hranici množiny přípustných řešení $\|\boldsymbol{\beta}\|_1 \leq t$. Červený bod značí hodnotu odhadu metodou obyčejných nejmenších čtverců. Modré elipsy odpovídají vrstevnicím funkce $\boldsymbol{\beta} \mapsto \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$. Červené tečkované polopřímky znázorňují zmenšovací členy pro $|\mathcal{A}| = 2$, tedy $\lambda(\mathbb{X}^\top\mathbb{X})^{-1}(1, 1)^\top$, respektive $\lambda(\mathbb{X}^\top\mathbb{X})^{-1}(-1, 1)^\top$, $\lambda \in [0, \infty)$. Fialová křivka odpovídá cestě odhadu $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$. Hodnoty konkrétních řešení metody lasso odpovídají bodům dotyku černého kosočtverce a modré elipsy.



Obrázek 2.3: Vyrovnané hodnoty odhadu metodou lasso. Fialová křivka odpovídá grafu funkce $\lambda \mapsto \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$. Červené body odpovídají vyrovnaným hodnotám odhadu metodou obyčejných nejmenších čtverců v modelu, který obsahuje vysvětlující proměnné X_1, resp. X_2, resp. (X_1, X_2). Modré polopřímky znázorňují rozdíl $\mathbb{X}_{\mathcal{M}}\widetilde{\boldsymbol{\beta}}_{\mathcal{M}}^{OLS} - \lambda\mathbb{X}_{\mathcal{M}}(\mathbb{X}_{\mathcal{M}}^\top\mathbb{X}_{\mathcal{M}})^{-1}\boldsymbol{s}_{\mathcal{M}}$, kde $\lambda\mathbb{X}_{\mathcal{M}}(\mathbb{X}_{\mathcal{M}}^\top\mathbb{X}_{\mathcal{M}})^{-1}\boldsymbol{s}_{\mathcal{M}}$ odpovídá vyrovnané hodnotě zmenšovacího členu metody lasso v modelu $\mathcal{M} \subseteq \{1, 2\}$. Zelené body odpovídají vyrovnaným hodnotám odhadů algoritmem LARS.

Nenormované sloupce matice $\mathbb{X}$ v (2.21) mají vliv na geometrickou interpretaci cesty odhadu, konkrétně na „ekvikorelační“ vektor $(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso})$. Pro $\lambda \in [0, \infty)$ je kosinus úhlu mezi vektory $\boldsymbol{X}_j$ a $\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ roven

$$\cos \alpha_j = \frac{\boldsymbol{X}_j^\top (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))}{\|\boldsymbol{X}_j\|_2 \cdot \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\|_2}$$

a pro aktivní složky odhadu plyne z (2.5)

$$|\cos \alpha_j| = \frac{\lambda}{\|\boldsymbol{X}_j\|_2 \cdot \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\|_2}, \qquad j \in \mathcal{A}(\lambda).$$

Za předpokladu normované odezvy je hodnota $\|\boldsymbol{X}_j\|_2 \cdot \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\|_2$ pro všechna $j \in \{1, \ldots, p\}$ stejná a speciálně absolutní hodnota $|\cos \alpha_j|$ je pro všechna $j \in \mathcal{A}(\lambda) = \operatorname{supp}(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))$ stejná. V případě různých euklidovských norem sloupců $\boldsymbol{X}_j$ a $\boldsymbol{X}_k$ je tato geometrická vlastnost porušena, neboť

$$|\cos \alpha_j| = \frac{|\lambda|}{\|\boldsymbol{X}_j\|_2 \cdot \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\|_2} \neq \frac{|\lambda|}{\|\boldsymbol{X}_k\|_2 \cdot \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\|_2} = |\cos \alpha_k|.$$

Pro normované vysvětlující proměnné nelze v případě $p = 2$ sestrojit data, pro která by cesta odhadu metodou lasso měla $(3^p + 1)/2$ uzlů – matice $\mathbb{X}^\top\mathbb{X}$ symetrická, se stejnými prvky na hlavní diagonále

$$\mathbb{X}^\top\mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_2 \end{pmatrix}.$$

Za předpokladu plné sloupcové hodnosti matice $\mathbb{X}$ je matice $(\mathbb{X}^\top\mathbb{X})^{-1}$ rovněž symetrická, se stejnými prvky na hlavní diagonále

$$(\mathbb{X}^\top\mathbb{X})^{-1} = \frac{1}{\det(\mathbb{X}^\top\mathbb{X})} \begin{pmatrix} \boldsymbol{X}_2^\top\boldsymbol{X}_2 & -\boldsymbol{X}_1^\top\boldsymbol{X}_2 \\ -\boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_1 \end{pmatrix}.$$

Symetrické matice se stejnými prvky na hlavní diagonále mají v $\mathbb{R}^2$ vždy vlastní vektory $(1, 1)^\top$ a $(1, -1)^\top$

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = (a+b) \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \begin{pmatrix} a & b \\ b & a \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = (a-b) \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad a, b \in \mathbb{R}.$$

Navíc skoro jistě platí $a = \|\boldsymbol{X}_1\|_2\|\boldsymbol{X}_2\|_2 > |\boldsymbol{X}_1^\top\boldsymbol{X}_2| = b$, čili $a \pm b > 0$, $a - b > 0$ a směr vlastního vektoru zůstane po aplikaci zobrazení určeného maticí $(\mathbb{X}^\top\mathbb{X})^{-1}$ stejný. Tedy v případě normovaných dat je zmenšovací člen $\lambda(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{s}$ ve směru $\boldsymbol{s}$ a vychází z $\widehat{\boldsymbol{\beta}}^{OLS}$. Dále například pro $\boldsymbol{s} = (1, 1)^\top$ je vektor $\boldsymbol{s}$ kolmý na stěny ℓ_1 koule, které leží v prvním a třetím kvadrantu a rovnoběžný se stěnami, které leží v druhém a čtvrtém kvadrantu. Aktivní složky řešení tedy musí být tvaru

$$\widehat{\boldsymbol{\beta}}^{lasso}_{\mathcal{A}(\lambda)}(\lambda) = \widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(\lambda)} - \lambda(\mathbb{X}^\top_{\mathcal{A}(\lambda)}\mathbb{X}_{\mathcal{A}(\lambda)})^{-1}\boldsymbol{s}_{\mathcal{A}(\lambda)} = \widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(\lambda)} - \lambda\gamma(\boldsymbol{s}_{\mathcal{A}(\lambda)})\boldsymbol{s}_{\mathcal{A}(\lambda)},$$

kde $\gamma(\boldsymbol{s}_{\mathcal{A}(\lambda)}) > 0$ je vlastní číslo matice $(\mathbb{X}^\top_{\mathcal{A}(\lambda)}\mathbb{X}_{\mathcal{A}(\lambda)})^{-1}$ příslušící vlastnímu vektoru $\boldsymbol{s}_{\mathcal{A}(\lambda)}(\lambda)$. Cesta odhadu má pro $p = 2$ v případě normovaných sloupců matice $\mathbb{X}$ vždy právě tři uzly a množinu řešení metody lasso (2.1) lze vyjádřit jako

$$\left\{ \begin{pmatrix} \operatorname{sign}(\widehat{\beta}_1^{OLS}) \cdot (|\widehat{\beta}_1^{OLS}| - \kappa)_+ \\ \operatorname{sign}(\widehat{\beta}_2^{OLS}) \cdot (|\widehat{\beta}_2^{OLS}| - \kappa)_+ \end{pmatrix} : \kappa \geq 0 \right\}.$$

Poznámka. V pozdější části prací jsme zjistili, že výše odvozená vlastnost je uvedena v Tibshirani (1996), sekce 2.3, druhý odstavec. Dále zmiňují, že pro $p > 2$ odhad metodou lasso již analogii této vlastnosti obecně nemá.

2.3.2 Bayesovský pohled

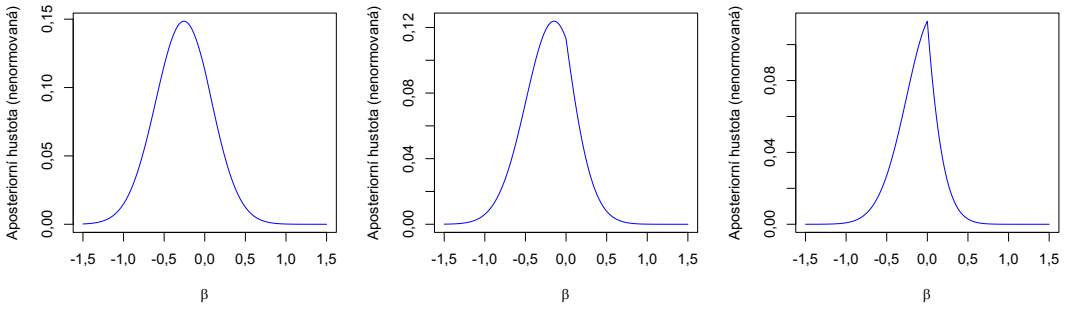
V Tibshirani (1996), sekce 5, poukazují na Bayesovskou interpretaci metody lasso. Předpokládejme normální lineární regresní model $Y \mid \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}^\top \boldsymbol{\beta}, \sigma^2)$, ve kterém jsou složky vektoru regresních koeficientů β_j, $j = 1, \ldots, p$, považovány za náhodný výběr z centrovaného Laplaceova rozdělení s parametrem škály $\theta > 0$. Na parametr θ lze nahlížet jako na apriorní přesvědčení o nulovosti daného regresního koeficientu. Apriorní rozdělení vektoru $\boldsymbol{\beta}$ má hustotu

$$\pi(\boldsymbol{\beta};\, \theta) = \prod_{j=1}^{p} \frac{\theta}{2} \exp\left\{-\theta|\beta_j|\right\}. \tag{2.22}$$

Označme $\kappa^{-1} = \sigma^2$. Na parametr $\kappa > 0$ lze nahlížet jako na apriorní důvěru v pozorovaná data. Aposteriorní hustota vektoru $\boldsymbol{\beta}$ je z Bayesovy věty tvaru ($\propto$ značí proporcionalitu)

$$\begin{aligned} f(\boldsymbol{\beta} \mid \boldsymbol{Y}, \mathbb{X};\, \kappa^{-1}, \theta) &\propto f(\boldsymbol{Y} \mid \boldsymbol{\beta}, \mathbb{X};\, \kappa^{-1}) \cdot \pi\left(\boldsymbol{\beta}\,;\, \theta\right) \\ &\propto \exp\left\{-\frac{\kappa}{2}\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 - \theta\|\boldsymbol{\beta}\|_1\right\}. \end{aligned} \tag{2.23}$$

Aposteriorní hustota (2.23) nabývá maxima v bodě $\widehat{\boldsymbol{\beta}}^{Lasso}(\theta/\kappa)$. Regularizační parametr λ v pojetí Bayesovské statistiky tedy odpovídá podílu $\lambda = \theta/\kappa$. Na obrázku 2.4 níže je vykreslena aposteriorní hustota (2.23), až na normalizační konstantu, pro $p = 1$ a různé hodnoty parametru λ. Poukažme zde na pozorování z Fan a Li (2001) – k produkci řídkých spojitých odhadů je nutné, aby penalizační funkce (respektive apriorní rozdělení $\boldsymbol{\beta}$) byla spojitá a měla singularitu v počátku.



(a) $\lambda = 0{,}1$ (b) $\lambda = 1$ (c) $\lambda = 3$

Obrázek 2.4: Aposteriorní hustota $f(\boldsymbol{\beta} \mid \boldsymbol{Y}, \mathbb{X}, \lambda) \propto \exp\{-\frac{1}{2}\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 - \lambda\|\boldsymbol{\beta}\|_1\}$, $\lambda = \theta/\kappa$, až na normalizační konstantu, $p = 1$, $n = 10$, $\boldsymbol{Y}^\top\boldsymbol{Y} = 4{,}36$, $\mathbb{X}^\top\boldsymbol{Y} = -2{,}23$, $\mathbb{X}^\top\mathbb{X} = 8{,}31$.

Parametry θ a κ jsou ve výše specifikovaném bayesovském modelu pevné, ale neznámé. Jako vhodné odhady se jeví odhad parametru Laplaceova rozdělení θ metodou maximální věrohodnosti $\widehat{\theta} = p/\|\boldsymbol{\beta}\|_1$ a odhad parametru normálního rozdělení σ^2 metodou maximální věrohodnosti $\widehat{\sigma}^2 = \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2/n$. Hodnoty $\boldsymbol{\beta}$ jsou ale rovněž neznámé, čili je třeba volit například $\widehat{\theta} = p/\|\widehat{\boldsymbol{\beta}}^{OLS}\|_1$ a $\widehat{\sigma}^2 = \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{OLS}\|_2^2/(n-p)$, což vede k odhadu parametru λ

$$\lambda^{Bayes} = \frac{p}{n-p}\frac{\|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{OLS}\|_2^2}{\|\boldsymbol{0} - \widehat{\boldsymbol{\beta}}^{OLS}\|_1}. \tag{2.24}$$

Takový odhad je ale možný pouze pro $n > p$. Nevýhodou odhadu (2.24) je, že neumožňuje kontrolovat počet vysvětlujících proměnných, které jsou odhadnuty jako nenulové. Výhodou odhadu (2.24) je jeho výpočetní nenáročnost. V následujícím příkladu porovnáme λ^{Bayes} s λ^{CV}, které minimalizuje chybu křížové validace.

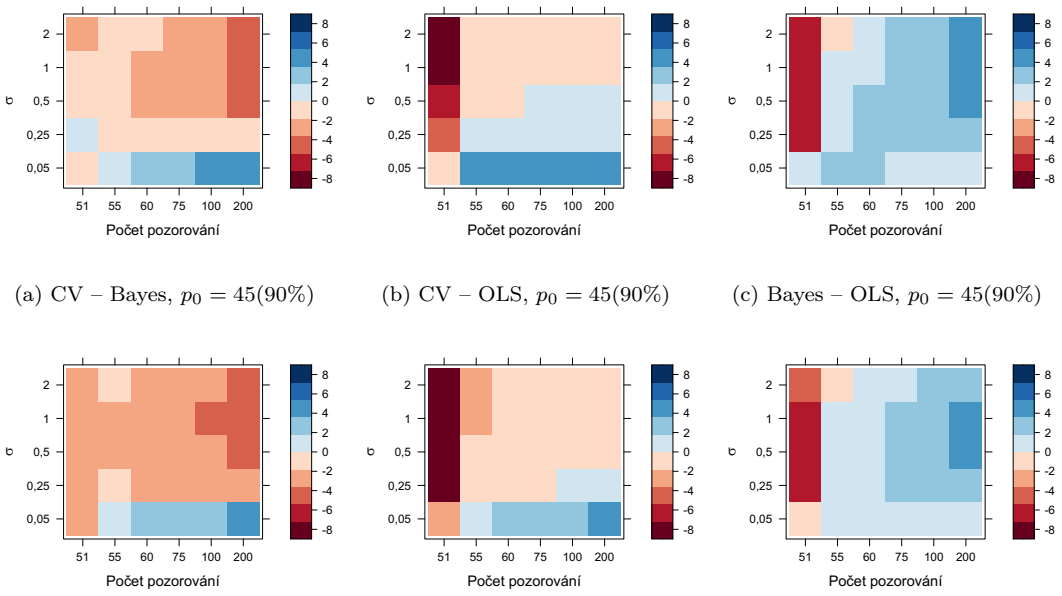
Příklad (Volba regularizačního parametru λ). Nechť $n, p \in \mathbb{N}$, $n > p$. Uvažujme vektor regresních koeficientů $\boldsymbol{\beta} \in \mathbb{R}^p$, jehož podmíněné rozdělení je tvaru

$$\psi(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma}) = \prod_{j=1}^{p} (\gamma_j \pi(\beta_j;\, \theta) + (1 - \gamma_j) \mathbb{I}[\beta_j = 0]),$$

kde $\pi(\beta_j;\, \theta)$ je hustota centrovaného Laplaceova rozdělení s parametrem $\theta > 0$ a $\gamma_j \in \{0,\, 1\}$, $j = 1, \ldots, p$, jsou parametry modelu. Nechť $\widetilde{\mathbb{X}} \in \mathbb{R}^{n \times p}$ je matice, jejíž složky tvoří náhodný výběr z rozdělení $\mathcal{N}(0,\, 1)$. Označme $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, kde $\boldsymbol{X}_j, j = 1, \ldots, p$, jsou studentizované sloupce matice $\widetilde{\mathbb{X}}$. Nechť

$$\widetilde{Y}_i = \sum_{j=1}^{p} X_{i,\,j} \beta_j + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,\, \sigma^2),\ i = 1, \ldots, n.$$

Označme $\boldsymbol{Y} = (\mathbb{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^\top) \widetilde{\boldsymbol{Y}}$. Pomocí simulace porovnáme odhad metodou lasso $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ s volbou regularizačního parametru: λ^{Bayes} vypočteného z (2.24), λ^{CV} které minimalizuje chybu křížové validace a $\lambda^{OLS} = 0$. Pro každou trojici hodnot $(p,\, n,\, \sigma)$, $p = 50$, $n \in \{p + 1,\, p + p/10,\, p + p/5,\, p + p/2,\, 2p,\, 4p\}$, $\sigma \in \{0{,}05,\, 0{,}25,\, 0{,}5,\, 1,\, 2\}$, byla 200-krát vygenerována data, vypočteny λ^{Bayes} a λ^{CV}, proveden odhad metodou lasso s danou volbou regularizačního parametru a vypočtena čtvercová chyba $\|\widehat{\boldsymbol{\beta}}^{lasso}(\lambda) - \boldsymbol{\beta}^*\|_2^2$. Pro každou trojici $(p,\, n,\, \sigma)$ byl vypočítán průměr z nasimulovaných čtvercových chyb. Výsledek simulace je shrnut na obrázku 2.5 níže.



(a) CV – Bayes, $p_0 = 45(90\%)$ (b) CV – OLS, $p_0 = 45(90\%)$ (c) Bayes – OLS, $p_0 = 45(90\%)$

(d) CV – Bayes, $p_0 = 25(50\%)$ (e) CV – OLS, $p_0 = 25(50\%)$ (f) Bayes – OLS, $p_0 = 25(50\%)$

Obrázek 2.5: Průměry z nasimulovaných hodnot $\log(\|\widehat{\boldsymbol{\beta}}^{lasso}(\lambda^A) - \boldsymbol{\beta}\|_2^2 / \|\widehat{\boldsymbol{\beta}}^{lasso}(\lambda^B) - \boldsymbol{\beta}\|_2^2)$ s volbou λ^{Bayes} pomocí vzorce (2.24), λ^{CV} pomocí křížové validace a $\lambda^{OLS} = 0$, pro různé hodnoty $\sigma \in \{0{,}05,\, 0{,}25,\, 0{,}5,\, 1,\, 2\}$ a $n \in \{p + 1,\, p + p/10,\, p + p/5,\, p + p/2,\, 2p,\, 4p\}$. Počet parametrů $p = 50$. První z dvojice odhadů λ v popisku obrázku dosahuje nižší čtvercové chyby $\|\widehat{\boldsymbol{\beta}}^{lasso}(\lambda) - \boldsymbol{\beta}\|_2^2$ pro $(n,\, \sigma)$ s červenými políčky. Například dle obrázku 2.5d vychází lépe volba $\lambda = \lambda^{CV}$, až na případ malého reziduálního rozptylu $\sigma^2 = 0{,}0025$ a $n \geq 55$.

Volba parametru λ pomocí křížové validace vychází nejlépe. Volba λ^{OLS} ve většině případů vede k lepším odhadům, než λ^{Bayes}. Případy, kdy je tomu naopak, přisuzujeme spíše přítomnosti regularizace, než konkrétní volbě λ^{Bayes}.

Poznámka. V pozdější části prací jsme zjistili, že v kontextu Bayesovského lasso navrhují v Park a Casella (2008) odhad obdobný (2.24) jako počáteční hodnotu parametru λ, která je v článku označena jako $\lambda^{(0)}$. Pro lasso jakožto metodu Bayesovské statistiky je v Park a Casella (2008) zavedena hierarchie

$$\begin{aligned} \boldsymbol{Y} \,|\, \mu,\, \mathbb{X},\, \boldsymbol{\beta},\, \sigma^2 &\sim \mathcal{N}_n(\mu \mathbf{1}_n + \mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbb{I}_n), \\ \boldsymbol{\beta} \,|\, \sigma^2,\, \tau_1^2, \ldots,\, \tau_p^2 &\sim \mathcal{N}_p(\mathbf{0}_p,\, \sigma^2 \mathbb{D}_{\boldsymbol{\tau}}), \quad \mathbb{D}_{\boldsymbol{\tau}} = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2), \\ \sigma^2,\, \tau_1^2, \ldots,\, \tau_p^2 &\sim \pi(\sigma^2) d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2, \quad \sigma^2,\, \tau_1^2, \ldots,\, \tau_p^2 > 0, \end{aligned} \tag{2.25}$$

a užit algoritmus, který v k-té iteraci pro parametr $\lambda^{(k-1)}$ provede *gibbsovo vzorkování* (Robert a Casella (1999), kapitola 9). K výpočtu parametru $\lambda^{(k)}$ je následně použit *Monte Carlo EM-algoritmus* Levine a Casella (2001) – Aposteriorní hustota je v proměnné λ maximalizována pro $\lambda^{(k)} = (2p/\sum_{j=1}^{p} \mathbb{E}_{\lambda^{(k-1)}}[\tau_j^2 \,|\, \boldsymbol{Y}])^{1/2}$, střední hodnota $\mathbb{E}_{\lambda^{(k-1)}}[\tau_j^2 \,|\, \boldsymbol{Y}]$ je odhadnuta průměrem z gibbsova vzorku.

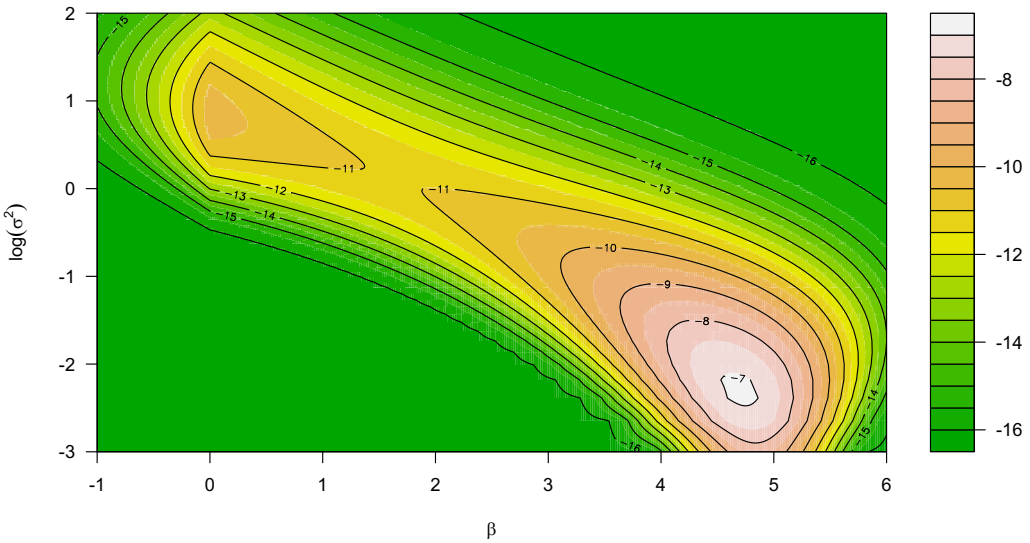
Hierarchie (2.25) využívá výsledku z Andrews a Mallows (1974) – Náhodnou veličinu s centrovaným Laplaceovým rozdělením lze vyjádřit jako podíl nezávislých náhodných veličin Z/V, kde Z má normované normální rozdělení a $(1/2)V^2$ má exponenciální rozdělení. Vektor regresních koeficientů $\boldsymbol{\beta}$ má tedy v modelu (2.25) podmíněnou apriorní hustotu

$$f(\boldsymbol{\beta} \,|\, \sigma^2;\, \lambda) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\{-\lambda |\beta_j| / \sqrt{\sigma^2}\}. \tag{2.26}$$

Dále dle Park a Casella (2008), budeme-li uvažovat hustotu σ^2 (až na normalizační konstantu) $\pi(\sigma^2) = 1/\sigma^2$ a místo (2.26) předpokládat (2.22), obdržíme sdružené aposteriorní rozdělení tvaru

$$f(\boldsymbol{\beta},\, \sigma^2 \,|\, \boldsymbol{Y},\, \mathbb{X},\, \theta) \propto (\sigma^2)^{-1-\frac{n-1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 - \theta \|\boldsymbol{\beta}\|_1 \right\}. \tag{2.27}$$

Sdružená aposteriorní hustota (2.27) může být bimodální. Na obrázku 2.6 níže uvádíme replikaci příkladu z Park a Casella (2008), Appendix B.



Obrázek 2.6: Transformace $\log(f(\beta,\, \sigma^2 \,|\, \boldsymbol{Y},\, \mathbb{X},\, \theta) + 10^{-7})$ hustoty (2.27) sdruženého aposteriorního rozdělení $(\boldsymbol{\beta},\, \sigma^2 \,|\, \boldsymbol{Y},\, \mathbb{X},\, \theta)$, až na normalizační konstantu.

2.4 Asymptotické vlastnosti

V této sekci se zabýváme limitním chováním odhadu metodou lasso. Důležitým výsledkem uvedeným v této sekci je, že odhad metodou lasso není konzistentním odhadem množiny aktivních složek vektoru regresních koeficientů.

Pro rozsah výběru $n \in \mathbb{N}$ označme $\lambda_n \geq 0$ příslušnou volbu regularizačního parametru a $\{\lambda_n\}_{n\in\mathbb{N}}$ posloupnost regularizačních parametrů. Posloupnost $\{\lambda_n\}_{n\in\mathbb{N}}$ budeme v následujících tvrzeních považovat za nenáhodnou. Níže uvádíme speciální verze tvrzení z Fu a Knight (2000) pro metodu lasso.

Poznámka. U metody lasso je běžnou volbou regularizačního parametru λ_n, které minimalizuje chybu křížové validace. Posloupnost $\{\lambda_n\}_{n\in\mathbb{N}}$ je v takovém případě náhodná, neboť závisí na náhodných datech. Analogie Věty 2.4, resp. Věty 2.5 platí i pro posloupnost náhodných veličin $\{\lambda_n\}_{n\in\mathbb{N}}$, využijeme-li věty o spojité transformaci, resp. Cramérovy-Slutského věty a budeme-li navíc předpokládat $\lambda_n/n \xrightarrow{\mathbb{P}} \lambda_0$, $n \to \infty$, resp. $\lambda_n/\sqrt{n} \xrightarrow{\mathbb{P}} \lambda_0$, $n \to \infty$, kde $\lambda_0 \geq 0$ je konstanta. Není ale zřejmé, zda jsou tyto předpoklady splněny pro posloupnost regularizačních parametrů $\{\lambda_n\}_{n\in\mathbb{N}}$, které minimalizují chybu křížové validace.

Věta 2.4 (Fu a Knight (2000), Theorem 1)**.** *Nechť* $\lim_{n\to\infty} \lambda_n/n = \lambda_0 \geq 0$ *a matice* $\lim_{n\to\infty} n^{-1}\mathbb{X}^\top\mathbb{X} = \mathbb{W}$ *je regulární, poté*

$$\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_n) \xrightarrow{\mathbb{P}} \operatorname*{argmin}_{\boldsymbol{\phi}\in\mathbb{R}^{\boldsymbol{p}}}(Z(\boldsymbol{\phi})), \qquad n \to \infty,$$

kde

$$Z(\boldsymbol{\phi}) = (\boldsymbol{\phi} - \boldsymbol{\beta}^*)^\top \mathbb{W}(\boldsymbol{\phi} - \boldsymbol{\beta}^*) + \lambda_0\|\boldsymbol{\phi}\|_1.$$

Věta 2.5 (Fu a Knight (2000), Theorem 2)**.** *Nechť* $\lim_{n\to\infty} \lambda_n/\sqrt{n} = \lambda_0 \geq 0$ *a matice* $\lim_{n\to\infty} n^{-1}\mathbb{X}^\top\mathbb{X} = \mathbb{W}$ *je regulární, poté*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_n) - \boldsymbol{\beta}^*) \xrightarrow{\mathcal{D}} \boldsymbol{u}^* = \operatorname*{argmin}_{u\in\mathbb{R}^p} V(\boldsymbol{u}), \qquad n \to \infty,$$

kde

$$V(\boldsymbol{u}) = -2\boldsymbol{u}^\top\boldsymbol{C} + \boldsymbol{u}^\top\mathbb{W}\boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} \Big(\operatorname{sign}(\beta_j^*)\mathbb{I}[\beta_j^* \neq 0]u_j + \mathbb{I}[\beta_j^* = 0]|u_j|\Big)$$

a $\boldsymbol{C} \sim \mathcal{N}_p(\mathbf{0}_p, \sigma^2\mathbb{W})$.

Tvrzení 2.6 (Zou (2006), Proposition 1)**.** *Je-li* $\lim_{n\to\infty} \lambda_n/\sqrt{n} = \lambda_0 \geq 0$, $p_{\mathcal{S}} < p$ *a matice* $\lim_{n\to\infty} n^{-1}\mathbb{X}^\top\mathbb{X} = \mathbb{W}$ *je regulární, poté*

$$\limsup_{n\to\infty} \mathbb{P}\left(\operatorname{supp}(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_n)) = \operatorname{supp}(\boldsymbol{\beta}^*)\right) \leq c < 1,$$

kde $c \in [0,\, 1)$ *je konstanta, jejíž hodnota závisí na skutečném modelu.*

Připomeňme značení $\mathcal{S} = \operatorname{supp}(\boldsymbol{\beta}^*)$ a $p_{\mathcal{S}} = |\mathcal{S}|$. Bez újmy na obecnosti budeme předpokládat $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_{p_{\mathcal{S}}}^*, 0, \ldots, 0)^\top$. Označme

$$\mathbb{W} = \begin{pmatrix} \mathbb{W}_{11} & \mathbb{W}_{12} \\ \mathbb{W}_{21} & \mathbb{W}_{22} \end{pmatrix},$$

kde $\mathbb{W}_{11} \in \mathbb{R}^{p_{\mathcal{S}}\times p_{\mathcal{S}}}$. Nyní uvedeme důkaz Tvrzení 2.6 z Zou (2006). Důkaz tvrzení 2.6 jsme si vybrali, neboť pracuje se složitě vyhlížejícím rozdělením náhodné veličiny $\boldsymbol{u}^*$ z Věty 2.5.

Důkaz. Zřejmě

$$\text{supp}(\widehat{\boldsymbol{\beta}}) = \mathcal{S} \Rightarrow \forall\, j \notin \mathcal{S} : \widehat{\beta}_j = 0$$

a tedy

$$\mathbb{P}(\text{supp}(\widehat{\boldsymbol{\beta}}) = \mathcal{S}) \leq \mathbb{P}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{-\mathcal{S}} = \mathbf{0}_{p-p_\mathcal{S}}).$$

Z věty 2.5, využitím věty o spojité transformaci, plyne $\sqrt{n}\widehat{\boldsymbol{\beta}}^{Lasso}_{-\mathcal{S}} \xrightarrow[n\to\infty]{\mathcal{D}} \boldsymbol{u}^*_{-\mathcal{S}}$. Z věty Portmanteauovy

$$\sqrt{n}\widehat{\boldsymbol{\beta}}^{Lasso}_{-\mathcal{S}} \xrightarrow[n\to\infty]{\mathcal{D}} \boldsymbol{u}^*_{-\mathcal{S}} \Rightarrow \limsup_{n\to\infty} \mathbb{P}(\sqrt{n}\widehat{\boldsymbol{\beta}}^{Lasso}_{-\mathcal{S}} = \mathbf{0}_{p-p_\mathcal{S}}) \leq \mathbb{P}(\boldsymbol{u}^*_{-\mathcal{S}} = \mathbf{0}_{p-p_\mathcal{S}}). \quad (2.28)$$

Nyní stačí volit $c = \mathbb{P}(\boldsymbol{u}^*_{-\mathcal{S}} = \mathbf{0}_{p-p_\mathcal{S}})$ a dokázat $\mathbb{P}(\boldsymbol{u}^*_{-\mathcal{S}} = \mathbf{0}_{p-p_\mathcal{S}}) < 1$.

Je-li $\lambda_0 = 0$, platí $\frac{\partial V(\boldsymbol{u})}{\partial \boldsymbol{u}} = -2\boldsymbol{C} + 2\mathbb{W}\boldsymbol{u}$. Z konvexity funkce $V(\boldsymbol{u})$ tedy

$$\boldsymbol{u}^* = \mathbb{W}^{-1}\boldsymbol{C} \sim \mathcal{N}_p(\mathbf{0}_p,\, \sigma^2\mathbb{W}^{-1}).$$

Rozdělení $\boldsymbol{u}^*_{-\mathcal{S}} \sim \mathcal{N}_{p-p_\mathcal{S}}(\mathbf{0}_{p-p_\mathcal{S}},\, \sigma^2\mathbb{W}^{-1}_{22})$ je spojité, čili $\mathbb{P}(\boldsymbol{u}^*_{-\mathcal{S}} = \mathbf{0}_{p-p_\mathcal{S}}) = 0 < 1$.

Je-li $\lambda_0 > 0$, odvodíme pravděpodobnost $\mathbb{P}(\boldsymbol{u}^*_{-\mathcal{S}} = \mathbf{0}_{p-p_\mathcal{S}})$ aplikací subgradientní podmínky optimality (1.8) na účelovou funkci V:

$$\mathbf{0}_p \in -2\boldsymbol{C} + 2\mathbb{W}\boldsymbol{u}^* + \lambda_0 \sum_{j=1}^{p} \left[\text{sign}(\beta^*_j)\mathbb{I}[\beta^*_j \neq 0] \cdot \boldsymbol{e}_j + \mathbb{I}[\beta^*_j = 0] \frac{\partial |u^*_j|}{\partial u^*_j} \cdot \boldsymbol{e}_j \right]. \quad (2.29)$$

Pro $j \in \mathcal{S}$ (tedy $\beta^*_j \neq 0$) a $u^*_j \in \mathbb{R}$ je podmínka (2.29) tvaru

$$0 = -2C_j + 2(\mathbb{W}\boldsymbol{u}^*)_j + \lambda_0 \text{sign}(\beta^*_j), \qquad j \in \mathcal{S}. \quad (2.30)$$

Pro $j \notin \mathcal{S}$ (tedy $\beta^*_j = 0$) a $u^*_j = 0$ je podmínka (2.29) tvaru

$$\lambda_0 \geq |-2C_j + 2(\mathbb{W}\boldsymbol{u}^*)_j|, \qquad j \notin \mathcal{S}. \quad (2.31)$$

Nechť $\boldsymbol{u}^*_{-\mathcal{S}} = \mathbf{0}_{p-p_\mathcal{S}}$, splnění podmínky (2.29) je poté ekvivalentní se splněním (2.30) a (2.31). Podmínka (2.30) je zřejmě ekvivalentní podmínce

$$\mathbf{0}_p = -2\boldsymbol{C}_\mathcal{S} + 2\mathbb{W}_{11}\boldsymbol{u}^*_\mathcal{S} + \lambda_0 \text{sign}(\boldsymbol{\beta}_\mathcal{S}) \quad (2.32)$$

a podmínka (2.31) je ekvivalentní s

$$\lambda_0 \geq |-2\boldsymbol{C}_{-\mathcal{S}} + 2\mathbb{W}_{21}\boldsymbol{u}^*_\mathcal{S}|, \quad (2.33)$$

kde nerovnost je míněna pro každou složku pravé strany. Vyjádřením $\boldsymbol{u}^*_\mathcal{S}$ v (2.32) a dosazením do (2.33) získáváme

$$\lambda_0 \geq |-2\boldsymbol{C}_{-\mathcal{S}} + \mathbb{W}_{21}\mathbb{W}^{-1}_{11}(2\boldsymbol{C}_\mathcal{S} - \lambda_0 \text{sign}(\boldsymbol{\beta}^*_\mathcal{S}))|.$$

Zřejmě

$$-2\boldsymbol{C}_{-\mathcal{S}} \sim \mathcal{N}_{p-p_\mathcal{S}}(\mathbf{0}_{p-p_\mathcal{S}},\, 4\sigma^2\mathbb{W}_{22}),$$

tedy

$$\mathbb{W}_{21}\mathbb{W}^{-1}_{11}(2\boldsymbol{C}_\mathcal{S} - \lambda_0 \text{sign}(\boldsymbol{\beta}^*_\mathcal{S})) \sim \mathcal{N}_{p-p_\mathcal{S}}(-\lambda_0\mathbb{W}_{21}\mathbb{W}^{-1}_{11}\text{sign}(\boldsymbol{\beta}^*_\mathcal{S}),\, \sigma^2\mathbb{W}_{21}\mathbb{W}^{-1}_{11}\mathbb{W}^\top_{21})$$

a

$$0 < c = \mathbb{P}(\lambda_0 \geq |-2\boldsymbol{C}_{-\mathcal{S}} + \mathbb{W}_{21}\mathbb{W}^{-1}_{11}(2\boldsymbol{C}_\mathcal{S} - \lambda_0 \text{sign}(\boldsymbol{\beta}^*_\mathcal{S}))|) < 1.$$

□

Poznámka. V Zou (2006), Appendix – Proof of Proposition 1, mají na levé straně implikace (2.28) v dolním indexu $\mathcal{S}$ (respektive v jejich značení $\mathcal{A}$, na prvním řádku strany 1426), věříme, že jde o přepis.

Značení (Landauova notace). Pro posloupnosti reálných čísel $\{\lambda_n\}_{n\in\mathbb{N}}$ a $\{r_n\}_{n\in\mathbb{N}}$ zavádíme následující značení

1. $\lambda_n = O(r_n)$ značí $\lim_{n\to\infty} |\lambda_n|/|r_n| \geq 0$,
2. $\lambda_n = o(r_n)$ značí $\lim_{n\to\infty} |\lambda_n|/|r_n| = 0$,
3. $\lambda_n = \omega(r_n)$ značí $\lim_{n\to\infty} |\lambda_n|/|r_n| = \infty$.

Dle Fu a Knight (2000) z vět 2.4 a 2.5 plyne, že konzistence odhadu $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_n)$ závisí na rychlosti růstu posloupnosti λ_n. Z věty 2.4 plyne, že pro $\lambda_n = O(n)$ je odhad metodou lasso konzistentní, ale vektor regresních koeficientů $\boldsymbol{\beta}^*$ je odhadem $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_n)$ identifikován pouze v případě $\lambda_n = o(n)$. Z věty 2.5 plyne, že pro $\sqrt{n}$-konzistenci odhadu je nutné $\lambda_n = O(\sqrt{n})$. Za předpokladu $\lambda_n = O(\sqrt{n})$ ale dle Zou (2006) z tvrzení 2.6 plyne, že množina aktivních složek odhadu je s kladnou pravděpodobností různá od množiny aktivních složek vektoru regresních koeficientů. Dle Fu a Knight (2000) je k „zajímavému“ asymptotickému rozdělení odhadu metodou lasso potřeba, aby posloupnost λ_n nerostla až příliš pomalu. Je-li $\lambda_n = o(\sqrt{n})$ pak z Věty 2.5 je limitní rozdělení odhadu metodou lasso tvaru $\boldsymbol{u}^* = \mathbb{W}^{-1}\boldsymbol{C} \sim \mathcal{N}_p(\boldsymbol{0}_p, \sigma^2\mathbb{W}^{-1})$, což je i limitním rozdělením odhadu metodou obyčejných nejmenších čtverců.

Odhad množiny aktivních složek metodou lasso v případě $\lambda_n = O(\sqrt{n})$ není konzistentní. V Zou (2006) se proto zabývali otázkou, zda lze konzistence odhadu množiny aktivních složek metodou lasso docílit v případech (1) $\lambda_n = \omega(n)$; (2) $\lim_{n\to\infty} \lambda_n/n = \lambda_0$, $0 < \lambda_0 < \infty$; nebo (3) $\lambda_n = o(n)\,\&\,\lambda_n = \omega(\sqrt{n})$. Ukazuje se, že ani to není možné. Věta 2.8 udává nutnou podmínku pro konzistenci odhadu množiny aktivních složek metodou lasso. Dříve ještě uvádíme lemma o asymptotickém rozdělení odhadu $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_n)$ v případě (3) $\lambda_n = o(n)\,\&\,\lambda_n = \omega(\sqrt{n})$.

Lemma 2.7 (Zou (2006), Lemma 3)**.** *Za předpokladu* $\lim_{n\to\infty} \lambda_n/\sqrt{n} = \infty$ *a zároveň* $\lim_{n\to\infty} \lambda_n/n = 0$ *platí*

$$\frac{\lambda_n}{n}(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_n) - \boldsymbol{\beta}^*) \xrightarrow{\mathcal{D}} \operatorname*{argmin}_{u\in\mathbb{R}^p} \widetilde{V}(\boldsymbol{u}), \qquad n \to \infty,$$

kde

$$\widetilde{V}(\boldsymbol{u}) = \boldsymbol{u}^\top \mathbb{W}\boldsymbol{u} + \sum_{j=1}^{p} \left(u_j \operatorname{sign}(\beta_j^*)\mathbb{I}[\beta_j \neq 0] + |u_j|\mathbb{I}[\beta_j^* = 0] \right).$$

Věta 2.8 (Zou (2006), Theorem 1)**.** *Nechť* $\lim_{n\to\infty} \mathbb{P}(\operatorname{supp}(\widehat{\boldsymbol{\beta}}^{Lasso}) = \mathcal{S}) = 1$, *poté existuje vektor znamének* $\boldsymbol{s} = (s_1, \ldots, s_{p_S})^\top$, $s_j \in \{-1, 1\}$ *takový, že*

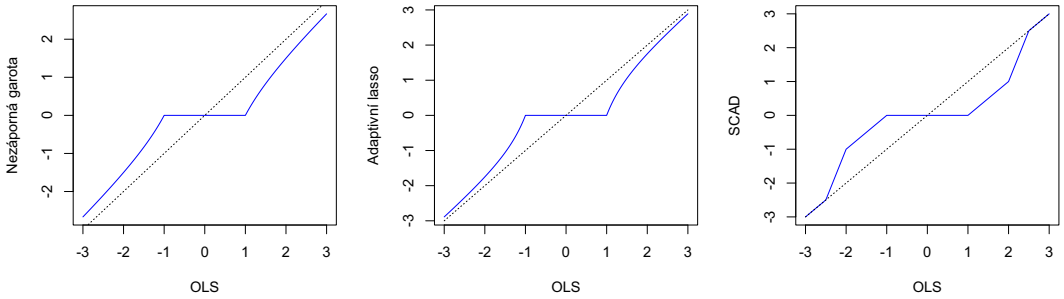
$$|\mathbb{W}_{21}\mathbb{W}_{11}^{-1}\boldsymbol{s}| \leq 1, \tag{2.34}$$

kde nerovnost je míněna pro každou složku levé strany.

V Zou (2006) dále dokazují, že podmínka (2.34) je netriviální. Ačkoli odhad metodou lasso nemá věštecké vlastnosti, modifikací účelové funkce (2.1) lze získat odhad, který věštecké vlastnosti má.

3. Adaptivní lasso

Odhad metodou lasso dle Tvrzení 2.6 s kladnou pravděpodobností neidentifikuje množinu aktivních složek vektoru regresních koeficientů $\mathcal{S}$. Existují ovšem metody, pomocí kterých již množinu $\mathcal{S}$ lze (asymptoticky) identifikovat. Mezi tyto metody patří *adaptivní lasso* Zou (2006), *SCAD* Fan a Li (2001) a *nezáporná garota* Breiman (1995). V této kapitole se zabýváme metodou adaptivní lasso.



(a) Nezáporná garota (b) Adaptivní lasso, $\gamma = 2$ (c) SCAD

Obrázek 3.1: Odhady metodami nezáporná garota, adaptivní lasso a SCAD, s volbou regularizačního parametru $\lambda = 1$, v případě ortogonální matice modelu, jako funkce odhadu metodou obyčejných nejmenších čtverců (OLS). Poznamenejme, že v případě ortogonální matice modelu je nezáporná garota speciálním případem adaptivního lasso s volbou parametru $\gamma = 1$.

3.1 Existence, tvar a jednoznačnost

Metodu adaptivní lasso lze považovat za zobecnění metody lasso, kde je každému z regresních koeficientů $\beta_1, \ldots, \beta_p$ přiřazena adaptivní váha.

Definice 3.1 (Odhad metodou adaptivní lasso)**.** *Nechť $\lambda \geq 0$, $\gamma > 0$, $\boldsymbol{Y} \in \mathbb{R}^n$ a $\mathbb{X} \in \mathbb{R}^{n \times p}$. Nechť $\widehat{\boldsymbol{\beta}}$ je $\sqrt{n}$-konzistentní odhad vektoru regresních koeficientů $\boldsymbol{\beta}^*$ a $\widehat{\boldsymbol{w}} = (1/|\widehat{\beta}_1|^\gamma, \ldots, 1/|\widehat{\beta}_p|^\gamma)$ je vektor adaptivních vah. Řešení úlohy*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j |\beta_j|, \qquad \lambda \geq 0. \tag{3.1}$$

budeme značit $\widehat{\boldsymbol{\beta}}^{AL}(\lambda; \widehat{\boldsymbol{w}})$ a nazývat odhad metodou adaptivní lasso.

Subgradientní podmínka optimality (1.8) pro úlohu (3.1) je tvaru

$$-\mathbb{X}^\top \boldsymbol{Y} + \mathbb{X}^\top \mathbb{X}\boldsymbol{\beta} = -\lambda \mathrm{diag}(\widehat{\boldsymbol{w}})\boldsymbol{s}, \qquad \boldsymbol{s} \in \partial \|\boldsymbol{\beta}\|_1, \tag{3.2}$$

kde $\boldsymbol{s} \in \partial \|\boldsymbol{\beta}\|_1$ je definováno jako v (2.2). Podmínku (3.2) lze přepsat jako

$$\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}} (\mathrm{diag}(\widehat{\boldsymbol{w}})\boldsymbol{\beta}) = \widetilde{\mathbb{X}}^\top \boldsymbol{Y} - \lambda \boldsymbol{s}, \qquad \boldsymbol{s} \in \partial \|\boldsymbol{\beta}\|_1, \tag{3.3}$$

kde $\widetilde{\mathbb{X}}^\top = \mathrm{diag}(\widehat{\boldsymbol{w}})^{-1} \mathbb{X}^\top$. Porovnáme-li (3.3) se subgradientní podmínkou optimality pro lasso (2.3), je patrné, že odhad metodou adaptivní lasso lze získat transformací odhadu metodou lasso jako $\widehat{\boldsymbol{\beta}}^{AL}(\lambda; \widehat{\boldsymbol{w}}) = \mathrm{diag}(\widehat{\boldsymbol{w}})^{-1} \boldsymbol{\beta}^{\star Lasso \star}(\lambda)$, kde

$$\boldsymbol{\beta}^{\star Lasso \star}(\lambda) \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\boldsymbol{Y} - \widetilde{\mathbb{X}}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1.$$

Vlastnosti odhadu metodou adaptivní lasso jsou tedy analogické vlastnostem odhadu metodou lasso z kapitoly 2, který je přenásobený maticí $\mathrm{diag}(\widehat{\boldsymbol{w}})^{-1}$.

3.2 Výpočetní algoritmus

Označme $\widetilde{\mathbb{X}} = \mathbb{X}\,\mathrm{diag}(\widehat{\boldsymbol{w}})^{-1}$. Pomocí modifikace algoritmu LARS pro lasso uvedeném v sekci 2.2 vypočteme předpis pro po částech afinní funkci

$$\boldsymbol{\beta}^{\star Lasso\star}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\mathrm{argmin}} \|\boldsymbol{Y} - \widetilde{\mathbb{X}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \qquad \lambda \geq 0.$$

Ze subgradientní podmínky optimality (3.3) plyne, že cesta odhadu metodou adaptivní lasso je tvaru $\{(\lambda,\, \mathrm{diag}(\widehat{\boldsymbol{w}})^{-1}\boldsymbol{\beta}^{\star Lasso\star}(\lambda)) : \lambda \geq 0\}$.

3.3 Geometrický a Bayesovský pohled

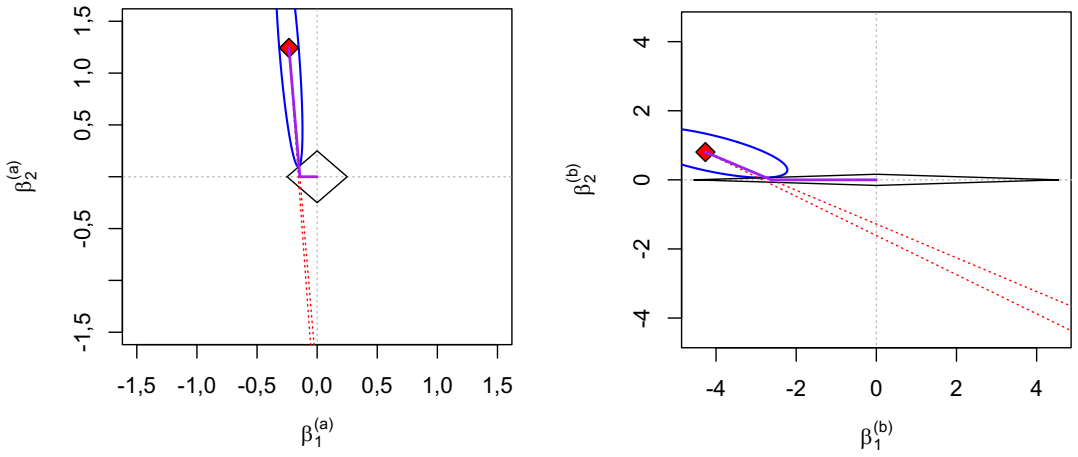
V této sekci uvádíme geometrický a Bayesovský pohled na metodu adaptivní lasso. Výše odvozený vztah $\widehat{\boldsymbol{\beta}}^{AL}(\lambda;\, \widehat{\boldsymbol{w}}) = \mathrm{diag}(\widehat{\boldsymbol{w}})^{-1}\boldsymbol{\beta}^{\star Lasso\star}(\lambda)$ znamená, že odhad metodou adaptivní lasso je ekvivalentní odhadu metodou lasso $\boldsymbol{\beta}^{\star Lasso\star}(\lambda)$, vyjádřeném v souřadnicích báze $\mathrm{diag}(\widehat{\boldsymbol{w}})$. Z Bayesovského pohledu je odhad metodou adaptivní lasso analogický odhadu metodou lasso s mírně komplexnější apriorní hustotou vektoru $\boldsymbol{\beta}$.

3.3.1 Geometrický pohled

Uvažujme následující data a odhad metodou obyčejných nejmenších čtverců

$$\boldsymbol{Y} = \begin{pmatrix} 2{,}00 \\ 0{,}50 \end{pmatrix}, \qquad \mathbb{X} = \begin{pmatrix} -0{,}63 & -0{,}84 \\ 0{,}18 & 1{,}59 \end{pmatrix}, \qquad \widehat{\boldsymbol{\beta}}^{OLS} = \begin{pmatrix} -4{,}27 \\ 0{,}80 \end{pmatrix}. \tag{3.4}$$

Pro parametr adaptivního lasso $\gamma = 2$ získáváme vektor vah $\widehat{\boldsymbol{w}} = (0{,}05,\, 1{,}55)^\top$. Inverzní váhy jsou $\widehat{\boldsymbol{w}}^{-1} = (18{,}23,\, 0{,}65)$. Na obrázku 3.2b je cesta odhadu metodou adaptivní lasso. Obrázek 3.2b lze získat z obrázku 3.2a transformací souřadnic $\beta_i^{(b)} = \widehat{w}_i^{-1}\beta_i^{(a)}$, $i \in \{1,\, 2\}$.



(a) $\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \widetilde{\mathbb{X}}\boldsymbol{\beta}\|_2^2$ s.t. $\|\boldsymbol{\beta}\|_1 \leq t$.

(b) $\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$ s.t. $\|\mathrm{diag}(\widehat{\boldsymbol{w}}^{-1})\boldsymbol{\beta}\|_1 \leq t$.

Obrázek 3.2: Na obrázku 3.2a je znázorněna cesta odhad metodou lasso pro transformovaná data $\widetilde{\mathbb{X}} = \mathbb{X}\,\mathrm{diag}(\widehat{\boldsymbol{w}})^{-1}$. Na obrázku 3.2b je cesta odhadu metodou adaptivní lasso, s volbou $\gamma = 2$. Červený bod značí hodnotu odhadu metodou obyčejných nejmenších čtverců pro příslušná data. Modré elipsy odpovídají vrstevnicím účelové funkce. Červené tečkované polopřímky znázorňují zmenšovací členy. Fialová křivka odpovídá cestě odhadu. Hodnoty odhadu pro konkrétní volbu $t \geq 0$ odpovídají bodům dotyku černého kosočtverce a modré elipsy.

3.3.2 Bayesovský pohled

Předpokládejme normální lineární regresní model $Y \mid \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}^\top \boldsymbol{\beta}, \kappa^{-1})$, ve kterém jsou složky vektoru regresních koeficientů β_j, $j = 1, \ldots, p$, nezávislé náhodné veličiny z centrovaného Laplaceova rozdělení s parametrem škály $\theta_j > 0$. Označme $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Apriorní hustota náhodného vektoru $\boldsymbol{\beta}$ je

$$\pi\left(\boldsymbol{\beta};\, \boldsymbol{\theta}\right) = \prod_{j=1}^{p} \frac{\theta_j}{2} \exp\left\{-\theta_j |\beta_j|\right\}.$$

Aposteriorní hustota je tvaru

$$f(\boldsymbol{\beta} \mid \boldsymbol{Y},\, \mathbb{X},\, \kappa^{-1},\, \boldsymbol{\theta}) \propto \exp\left\{-\frac{\kappa}{2}\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 - \sum_{j=1}^{p} \theta_j |\beta_j|\right\}. \tag{3.5}$$

Hustota $f(\boldsymbol{\beta} \mid \boldsymbol{Y},\, \mathbb{X},\, \kappa^{-1},\, \boldsymbol{\theta})$ nabývá maxima v bodě

$$\operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\,\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \kappa^{-1} \sum_{j=1}^{p} \theta_j |\beta_j|. \tag{3.6}$$

Řešením (3.6) je $\widehat{\boldsymbol{\beta}}^{AL}(\kappa^{-1},\, \boldsymbol{\theta})$. Odhad parametru θ_j maximalizující apriorní hustotu $\pi(\boldsymbol{\beta};\, \boldsymbol{\theta})$ v argumentu θ_j je $\widehat{\theta}_j = 1/|\beta_j|$. Jelikož hodnoty náhodného vektoru $\boldsymbol{\beta}$ nepozorujeme, volí se $\widehat{\theta}_j = 1/|\widehat{\beta}_j|$, kde $\widehat{\boldsymbol{\beta}}$ je nějaký $\sqrt{n}$-konzistentní odhad parametru $\boldsymbol{\beta}$, například metodou nejmenších čtverců nebo hřebenovou regresí.

Obecnější Bayesovský pohled

Lasso i adaptivní lasso lze zařadit do širší třídy Bayesovských směsových modelů. Regresní koeficienty $\beta_1, \ldots, \beta_p$, považujeme za nezávislé náhodné veličiny a regresní koeficient β_j, $j \in \{1, \ldots, p\}$, považujeme za náhodnou veličinu z centrovaného Laplaceova rozdělení s parametrem škály $\theta_j \geq 0$. Předpokládejme nyní, že θ_j je náhodná veličina s distribuční funkcí $\Psi(\theta)$, čili marginální rozdělení náhodné veličiny β_j je tvaru

$$\pi(\beta_j) = \int \pi(\beta_j \mid \theta_j) d\Psi(\theta_j).$$

Metoda lasso odpovídá volbě $\Psi(\theta) = \mathbb{I}[\theta \leq \theta^*]$, kde $\theta^* \geq 0$ je zvolená hodnota. Metoda adaptivní lasso v naší zjednodušené hierarchii odpovídá volbě $\psi(\theta_j) \propto 1$, tedy že $\theta_1, \ldots, \theta_p$ jsou náhodné veličiny ze spojitého neiformativního rovnoměrného rozdělení na $(-\infty, \infty)$. Výhodou metod lasso a adaptivní lasso je, že umožňují snadný přechod mezi Bayesovskou a frekventistickou formulací.

V Bayesovské statistice je dalším vhodným modelem pro řídká data takzvané *spike-and-slab lasso* Ročková a George (2018) s apriorním rozdělením

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \prod_{j=1}^{p} [\gamma_j \pi(\beta_j \mid \lambda_1) + (1 - \gamma_j)\pi(\beta_j \mid \lambda_0)], \qquad \gamma_j \in \{0,\, 1\}, \tag{3.7}$$

kde λ_0 volíme velké, čili hustota $\pi(\beta;\, \lambda_0)$ je „špičatá“, λ_1 volíme malé, čili hustota $\pi(\beta;\, \lambda_1)$ je „placatá“ a $\gamma_1, \ldots, \gamma_p$ je náhodný výběr z alternativního rozdělení. V následující sekci uvádíme asymptotické vlastnosti metody adaptivní lasso.

3.4 Asymptotické vlastnosti

Dle Věty 3.2 vhodná volba posloupnosti regularizačních parametru $\{\lambda_n\}_{n\in\mathbb{N}}$ a adaptivních vah $\widehat{\boldsymbol{w}} = (1/|\widehat{\beta}_1|^\gamma, \ldots, 1/|\widehat{\beta}_p|^\gamma)$ vede k věšteckým vlastnostem odhadu metodou adaptivní lasso.

Věta 3.2 (Zou (2006), Theorem 2)**.** *Nechť* $\lambda_n = o(\sqrt{n})$, $\lim_{n\to\infty} \lambda_n n^{(\gamma-1)/2} = \infty$, *matice* $\lim_{n\to\infty} n^{-1}\mathbb{X}^\top\mathbb{X} = \mathbb{W}$ *je regulární,* $\widehat{\boldsymbol{\beta}}$ *je* $\sqrt{n}$*-konzistentní odhad vektoru regresních koeficientů,* $\gamma > 0$ *a* $\widehat{\boldsymbol{w}} = (1/|\widehat{\beta}_1|^\gamma, \ldots, 1/|\widehat{\beta}_p|^\gamma)$. *Poté má odhad metodou adaptivní lasso* $\widehat{\boldsymbol{\beta}}^{AL}(\lambda_n; \widehat{\boldsymbol{w}})$ *věštecké vlastnosti, tedy:*

1. *Odhad asymptoticky přesně identifikuje množinu aktivních složek vektoru regresních koeficientů:* $\lim_{n\to\infty} \mathbb{P}(\mathrm{supp}(\widehat{\boldsymbol{\beta}}^{AL}) = \mathrm{supp}(\boldsymbol{\beta}^*)) = 1$.

2. *Odhad* $\widehat{\boldsymbol{\beta}}^{AL}_{\mathcal{S}}$ *je asymptoticky normální:* $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{AL}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_{p_{\mathcal{S}}}(0, \sigma^2\mathbb{W}^{-1}_{11})$.

Pro $j \in \mathrm{supp}(\boldsymbol{\beta}^*)$ vede předpoklad $\lambda_n = o(\sqrt{n})$ společně s konzistencí $\widehat{\boldsymbol{\beta}}$ (konzistence $\widehat{\boldsymbol{\beta}}$ je důsledkem $\sqrt{n}$-konzistence $\widehat{\boldsymbol{\beta}}$) k asymptotické normalitě odhadu metodou adaptivní lasso. Tento výsledek je analogií výsledku z Věty 2.5. Pro přesný odhad regresních koeficientů β^*_j, $j \notin \mathrm{supp}(\boldsymbol{\beta}^*)$, je k předpokladům Věty 2.5 přidat předpoklady $\lim_{n\to\infty} \lambda_n n^{(\gamma-1)/2} = \infty$ a $\sqrt{n}$-konzistence odhadu $\widehat{\boldsymbol{\beta}}$, které zajišťují, že příslušný penalizační člen adaptivního lasso $\lambda_n \widehat{w}_j = \lambda_n/|\widehat{\beta}_j|^\gamma$ bude divergovat do nekonečna pro $n \to \infty$. Předpoklad existence $\sqrt{n}$-konzistentního odhadu $\widehat{\boldsymbol{\beta}}$ vektoru regresních koeficientů není v porovnání s předpoklady metody lasso silný, neboť z Fu a Knight (2000), Theorem 2, plyne, že $\sqrt{n}$-konzistentní odhad vektoru regresních koeficientů lze získat hřebenovou regresí, s volbou posloupnosti regularizačních parametrů $\nu_n = O(\sqrt{n})$. Z (1.20) je odhad hřebenovou regresí (pro vhodné $\nu_n \geq 0$) jednoznačně definován pro libovolné $n \in \mathbb{N}$. Je ovšem třeba brát v potaz, že výsledky Věty 3.2 jsou asymptotické a zejména pro malý počet pozorování nemusí být volba vah $\widehat{\boldsymbol{w}} = (1/|\widehat{\beta}_1|, \ldots, 1/|\widehat{\beta}_p|)$ spolehlivá a odhad metodou adaptivní lasso nemusí být lepší, než odhad metodou lasso.

Na regularizované odhady není vhodné aplikovat nám doposud známé schéma odhad—asymptotické vlastnosti—asymptotické intervaly spolehlivosti, ani nám doposud známý přístup k testování podmodelů neboť:

a) Jak jsme ukázali v sekci 2.4, regularizované odhady díky vychýlení nemusí identifikovat předmět našeho zájmu, v našem případě vektor regresních koeficientů $\boldsymbol{\beta}^{\mathcal{M}}$, kde $\mathcal{M} \subseteq \{1, \ldots, p\}$ je zvolený model.

b) Asymptotické rozdělení regularizovaného odhadu může snadno vést k antikonzervativním intervalovým odhadům. Je-li například $\widehat{\beta}^{AL}_j = 0$, příslušný intervalový odhad na základě asymptotického rozdělení z Věty 3.2 je $[0, 0]$, přičemž $\mathbb{P}([0, 0] \ni \beta^*_j) \in \{0, 1\}$.

c) Připomeňme značení $\mathcal{A}(\lambda)$ množiny aktivních složek odhadu metodou lasso a $\lambda_k \geq \lambda_{k+1} \geq 0$ uzlů modifikace algoritmu LARS. Nechť $\mathcal{A}(\lambda_k) \subseteq \mathcal{A}(\lambda_{k+1})$. Podmodel $\mathcal{A}(\lambda_k)$ modelu $\mathcal{A}(\lambda_{k+1})$ je zvolen v jistém smyslu optimálně a klasický F-test na podmodel může vést k antikonzervativním výsledkům.

V následující kapitole proto uvedeme dvě metody statistické inference. První z uvedených metod řeší problém konstrukce intervalových odhadů pro vektor regresních koeficientů $\boldsymbol{\beta}^{\mathcal{A}(\lambda)}$. Druhá z uvedených metod je analogií F-testu pro pár modelů $\mathcal{A}(\lambda_k) \subseteq \mathcal{A}(\lambda_{k+1})$.

4. Statistická inference pro lasso

V úvodu této kapitoly popíšeme problematiku statistické inference po výběru rysů Berk a kol. (2013), Leeb a Pötscher (2005) na základě množiny aktivních složek řídkého odhadu. V sekci 4.2 je uvedena metoda ke konstrukci intervalových odhadů v normálním lineárním regresním modelu, jehož vysvětlující proměnné byly vybrány na základě množiny aktivních složek odhadu metodou lasso Lee a kol. (2016). V sekci 4.3 je uveden *kovarianční test* Lockhart a kol. (2014), který testuje signifikanci složky vstupující do množiny aktivních složek odhadu metodou lasso.

4.1 Statistická inference po výběru rysů

Metody produkující řídké odhady lze využít k výběru rysů. Uvažujme metodu produkující řídké odhady v lineárním regresním modelu. Označme $\mathcal{A}$ množinu aktivních složek tohoto řídkého odhadu. Složky neznámého vektoru regresních koeficientů $\boldsymbol{\beta}^*$ jsou odhadem identifikovány jako signifikantní v případě $j \in \mathcal{A}$ a identifikovány jako insignifikantní v případě $j \notin \mathcal{A}$. Jeví se proto možné provést výběr modelu $\mathcal{M}$, který bude použit k statistické analýze na základě množiny aktivních složek odhadu $\mathcal{A}$, tedy k analýze volíme odhadem identifikovaný model $\mathcal{M} = \mathcal{A}$. Tento přístup s sebou ale nese jisté problémy.

Klasický přístup k testování hypotéz předpokládá, že jevy v nulové hypotéze i alternativě byly předem pevně zvoleny, nezávisle na pozorovaných datech. Klasický přístup ke konstrukci intervalových odhadů v lineární regresním modelu předpokládá, že model byl předem pevně zvolen. Množina aktivních složek odhadu $\mathcal{A}$ je funkcí náhodných vstupních dat $\boldsymbol{Y}$, $\mathbb{X}$ a je zvolena v jistém ohledu optimálně. U následné inference v modelu $\mathcal{M} = \mathcal{A}$ poté hrozí vyšší pravděpodobnost chyby prvního druhu, než je předem stanovená hladina $\alpha \in (0, 1)$. Tento problém ilustrujeme příkladem z Lockhart a kol. (2014).

Příklad. Regrese postupným krokem vpřed[1] je iterační algoritmus, který provádí automatický výběr rysů. Algoritmus začíná s prázdnou množinou aktivních složek $\mathcal{A}(0) := \emptyset$ a v jednotlivých krocích do množiny aktivních složek přidává indexy, které vedou k optimalizaci zvoleného kritéria. Nechť pro $k \in \{0, \ldots, p-1\}$ je v $(k+1)$-ním kroku algoritmu do množiny aktivních složek $\mathcal{A}(k)$ přidán index složky, která maximalizuje pokles reziduálního součtu čtverců a zároveň doposud není v $\mathcal{A}(k)$. Tedy $\mathcal{A}(k+1) = \mathcal{A}(k) \cup \{j\}$, kde

$$j = \underset{j \in \{1,\ldots,p\} \setminus \mathcal{A}(k)}{\operatorname{argmax}} \left\{ \|\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(k)} \widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(k)}\|_2^2 - \|\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(k)\cup\{j\}} \widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(k)\cup\{j\}}\|_2^2 \right\}$$

a $\widetilde{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}(k)} = (\mathbb{X}^\top_{\mathcal{A}(k)} \mathbb{X}_{\mathcal{A}(k)})^{-1} \mathbb{X}^\top_{\mathcal{A}(k)} \boldsymbol{Y}$ je odhad vektoru regresních koeficientů metodou obyčejných nejmenších čtverců v modelu $\mathcal{A}(k)$. Metoda v jednotlivých krocích přidává index, který maximalizuje hodnotu čitatele testové statistiky F-testu pro pár modelů $\mathcal{A}(k) \subset \mathcal{A}(k+1)$. Testová statistika $F_{\mathcal{A}(k),\,\mathcal{A}(k+1)}$ tedy bude mít tendenci nabývat vyšších hodnot, než za předpokladu předem pevně zvoleného páru model–podmodel a výsledný test bude antikonzervativní, čili pravděpodobnost chyby prvního druhu bude vyšší, než předem stanovená hladina $\alpha \in (0, 1)$.

[1] Anglicky: *forward stepwise regression*

Jako možné řešení výše popsaného problému se jeví odvodit rozdělení testové statistiky podmíněně vybraným modelem. Například v případě uvedeného příkladu by k testování podmodelů bylo vhodné podmínit alespoň náhodným jevem $\{\mathcal{M} = \mathcal{A}(k+1)\}$, čili použít testovou statistiku $F_{\mathcal{A}(k),\mathcal{M}} \,|\, \mathcal{M} = \mathcal{A}(k+1)$. Ideu podmíněné statistické inference popíšeme na konstrukci intervalů spolehlivosti. Uvažujme regresní koeficient β_j, $j \in \{1, \ldots, p\}$, pro který chceme sestavit intervalový odhad. Nechť $\mathcal{A}$ je model zvolený na základě náhodných vstupních dat a $\boldsymbol{\beta}^{\mathcal{A}}$ je vektor regresních koeficientů v modelu $\mathcal{A}$. Jestliže $j \notin \mathcal{A}$, poté regresní koeficient $\beta_j^{\mathcal{A}}$ není v modelu $\mathcal{A}$ definován a nelze pro něj sestrojit intervalový odhad $\boldsymbol{C}_j^{\mathcal{A}}$. Čili náhodný jev $\{\boldsymbol{C}_j^{\mathcal{A}} \ni \beta_j^{\mathcal{A}}\}$ není dobře definován a intervalové odhady nelze konstruovat na základě kritéria

$$\mathbb{P}(\boldsymbol{C}_j^{\mathcal{A}} \ni \beta_j^{\mathcal{A}}) \geq 1 - \alpha. \tag{4.1}$$

Jev $\{\boldsymbol{C}_j^{\mathcal{A}} \ni \beta_j^{\mathcal{A}}\}$ je definován pouze pro modely, pro které platí $j \in \mathcal{A}$. Dle Berk a kol. (2013), sekce 4.3, by tedy bylo třeba kontrolovat pravděpodobnost pokrytí

$$\mathbb{P}(\boldsymbol{C}_j^{\mathcal{A}} \ni \beta_j^{\mathcal{A}} \,|\, j \in \mathcal{A}) \geq 1 - \alpha. \tag{4.2}$$

Jelikož se statistická inference provádí pouze na vysvětlujících proměnných, které jsou obsaženy ve zvoleném modelu $\mathcal{A}$, měla by kontrola podmíněné pravděpodobnosti $\mathbb{P}(\boldsymbol{C}_j^{\mathcal{A}} \ni \beta_j^{\mathcal{A}} \,|\, j \in \mathcal{A})$ být dostatečná. Postačující podmínkou k (4.2) je

$$\forall \mathcal{M} \in 2^{\{1,\ldots,p\}} \, \forall j \in \mathcal{M} : \mathbb{P}(\boldsymbol{C}_j^{\mathcal{M}} \ni \beta_j^{\mathcal{M}} \,|\, \mathcal{A} = \mathcal{M}) \geq 1 - \alpha, \tag{4.3}$$

kde $2^{\{1,\ldots,p\}}$ je potenční množina množiny $\{1, \ldots, p\}$, neboť využitím věty o úplné pravděpodobnosti na (4.2) získáváme

$$\mathbb{P}(\boldsymbol{C}_j^{\mathcal{A}} \ni \beta_j^{\mathcal{A}} \,|\, j \in \mathcal{A}) = \sum\nolimits_{\substack{\mathcal{M} \in 2^{\{1,\ldots,p\}} \\ \mathcal{M} : j \in \mathcal{M}}} \mathbb{P}(\boldsymbol{C}_j^{\mathcal{M}} \ni \beta_j^{\mathcal{M}} \,|\, \mathcal{A} = \mathcal{M}) \cdot \mathbb{P}(\mathcal{A} = \mathcal{M} \,|\, j \in \mathcal{M}).$$

Pro mnohonásobné porovnávání navrhují v Berk a kol. (2013) kontrolu *chyby na úrovni modelu*[2]

$$\forall \mathcal{M} \in 2^{\{1,\ldots,p\}} : \mathrm{FWER} = \mathbb{P}(\beta_j^{\mathcal{M}} \notin \boldsymbol{C}_j^{\mathcal{M}},\, j \in \mathcal{M}) \leq \alpha.$$

Dle Lee a kol. (2016) je pro modely s vysokým počtem parametrů p toto kritérium příliš striktní a místo toho navrhují kontrolovat *falešnou míru pokrytí*[3]

$$\mathrm{FCR} = \mathbb{E}\left[\frac{|\{j \in \mathcal{A} : \beta_j^{\mathcal{A}} \notin \boldsymbol{C}_j^{\mathcal{A}}\}|}{|\mathcal{A}|};\, |\mathcal{A}| > 0\right] \leq \alpha. \tag{4.4}$$

Postačující podmínkou pro (4.4) je dle Lee a kol. (2016), Lemma 2.1, podmínka (4.3), neboť využitím (4.3) a věty o úplně pravděpodobnosti získáváme

$$\begin{aligned}
\mathrm{FCR} &= \sum\nolimits_{\substack{\mathcal{M} \in 2^{\{1,\ldots,p\}} \\ |\mathcal{M}|>0}} \sum_{j \in \mathcal{M}} \frac{\mathbb{P}(\beta_j^{\mathcal{M}} \notin \boldsymbol{C}_j^{\mathcal{M}} \,|\, \mathcal{A} = \mathcal{M})}{|\mathcal{M}|} \cdot \mathbb{P}(\mathcal{A} = \mathcal{M};\, |\mathcal{A}| > 0) \\
&\leq \sum\nolimits_{\substack{\mathcal{M} \in 2^{\{1,\ldots,p\}} \\ |\mathcal{M}|>0}} \frac{|\mathcal{M}|\alpha}{|\mathcal{M}|} \cdot \mathbb{P}(\mathcal{A} = \mathcal{M} \,|\, |\mathcal{A}| > 0) = \alpha.
\end{aligned}$$

V následující sekci odvodíme pravděpodobnostní rozdělení náhodné veličiny $(\widetilde{\beta}_{j,\mathcal{M}}^{OLS} \,|\, \mathcal{M} = \mathcal{A}(\lambda))$, kde $\widetilde{\beta}_{j,\mathcal{M}}^{OLS}$ značí odhad regresního koeficientu $\beta_j^{\mathcal{M}}$ metodou obyčejných nejmenších čtverců v modelu $\mathcal{M}$.

[2] Anglicky: *familywise error rate*

[3] Anglicky: *false coverage rate*

4.2 Přesná inference po výběru rysů

Dle Lee a kol. (2016) lze pro množinu aktivních složek odhadu metodou lasso $\mathcal{A}(\lambda)$, $\lambda \geq 0$, odvodit rozdělení náhodného jevu $\{\mathcal{A}(\lambda) = \mathcal{M}\}$, $\mathcal{M} \in 2^{\{1,\dots,p\}}$, díky čemuž je za předpokladu $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{\mu} \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n}$, možné odvodit rozdělení náhodné veličiny $(\boldsymbol{\eta}^\top \boldsymbol{Y} \,|\, \mathcal{A}(\lambda) = \mathcal{M})$, kde $\boldsymbol{\eta} \in \mathbb{R}^n$ je zvolený směr. Ke konstrukci intervalů spolehlivosti pro odhad regresního koeficientu β_j metodu obyčejných nejmenších čtverců volíme směr $\boldsymbol{\eta} = \boldsymbol{e}_j(\mathbb{X}_{\mathcal{M}}^\top \mathbb{X}_{\mathcal{M}})^{-1}\mathbb{X}_{\mathcal{M}}^\top$, kde $\mathcal{M} \subseteq \{1,\dots,p\}$ je zvolený model. Z rozdělení $(\boldsymbol{\eta}^\top \boldsymbol{Y} \,|\, \mathcal{A}(\lambda) = \mathcal{M})$ jsme schopni sestrojit intervaly spolehlivosti s vlastností (4.3).

V této sekci budeme po vzoru Lee a kol. (2016) předpokládat normální rozdělení odezvy $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$. Připomeňme značení znamének odhadu metodou lasso $\boldsymbol{s}(\lambda) = \operatorname{sign}(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda))$. Dle Lee a kol. (2016), Corollary 4.4, lze jev $\{\mathcal{A}(\lambda) = \mathcal{M}\}$ rozložit na sjednocení konvexních vazebních podmínek na náhodnou veličinu $\boldsymbol{Y}$. Toto sjednocení je tvaru

$$\begin{aligned}\{\mathcal{A}(\lambda) = \mathcal{M}\} &= \bigcup\nolimits_{\boldsymbol{s}\in\{-1,1\}^{|\mathcal{M}|}} \{\mathcal{A}(\lambda) = \mathcal{M},\, \boldsymbol{s}(\lambda) = \boldsymbol{s}\} \\ &= \bigcup\nolimits_{\boldsymbol{s}\in\{-1,1\}^{|\mathcal{M}|}} \{\mathbb{A}(\mathcal{M},\, \boldsymbol{s})\boldsymbol{Y} < \boldsymbol{b}(\mathcal{M},\, \boldsymbol{s})\},\end{aligned} \tag{4.5}$$

kde $\boldsymbol{s} \in \{-1,\, 1\}^{|\mathcal{M}|}$ je vektor znamének a matice $\mathbb{A}(\mathcal{M},\, \boldsymbol{s}) \in \mathbb{R}^{q\times n}$ a vektor $\boldsymbol{b}(\mathcal{M},\, \boldsymbol{s}) \in \mathbb{R}^q$, $q \in \mathbb{N}$, jsou dány vztahy z Lee a kol. (2016), Proposition 4.2. Množina řešení soustavy lineárních nerovnic $\{\boldsymbol{y} \in \mathbb{R}^n : \mathbb{A}(\mathcal{M},\, \boldsymbol{s})\boldsymbol{y} < \boldsymbol{b}(\mathcal{M},\, \boldsymbol{s})\}$ je otevřený konvexní polyedr, neboť se jedná o průnik otevřených poloprostorů. Rozdělení náhodné veličiny $(\boldsymbol{\eta}^\top \boldsymbol{Y} \,|\, \{\mathcal{A}(\lambda) = \mathcal{M}\})$ odvodíme pomocí rozdělení náhodných veličin $(\boldsymbol{\eta}^\top \boldsymbol{Y} \,|\, \{\mathcal{A}(\lambda) = \mathcal{M},\, \boldsymbol{s}_{\mathcal{A}(\lambda)} = \boldsymbol{s}\})$, $\mathcal{M} \in 2^{\{1,\dots,p\}}$, $\boldsymbol{s} \in \{-1,\, 1\}^{|\mathcal{M}|}$.

4.2.1 Podmiňování jedním polyedrem

Nechť $q \in \mathbb{N}$, $\mathbb{A} \in \mathbb{R}^{n\times q}$ je matice a $\boldsymbol{\eta} \in \mathbb{R}^n$ a $\boldsymbol{b} \in \mathbb{R}^q$ jsou vektory. K odvození rozdělení $\boldsymbol{\eta}^\top \boldsymbol{Y} | \{\mathbb{A}\boldsymbol{Y} < \boldsymbol{b}\}$ využijeme následující lemma.

Lemma 4.1 (Lee a kol. (2016), Lemma 5.1)**.** *Nechť $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{\eta} \in \mathbb{R}^n$, $\boldsymbol{c} = \Sigma\boldsymbol{\eta}(\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta})^{-1}$ a $\boldsymbol{Z} \equiv (\mathbb{I}_n - \boldsymbol{c}\boldsymbol{\eta}^\top)\boldsymbol{Y}$. Poté*

$$\{\mathbb{A}\boldsymbol{Y} < \boldsymbol{b}\} = \{\mathcal{V}^-(\boldsymbol{Z}) \leq \boldsymbol{\eta}^\top\boldsymbol{Y} \leq \mathcal{V}^+(\boldsymbol{Z}),\, \mathcal{V}^0(\boldsymbol{Z}) \geq 0\},$$

kde

$$\mathcal{V}^-(\boldsymbol{Z}) \equiv \max_{j:(\mathbb{A}\boldsymbol{c})_j<0} \frac{b_j - (\mathbb{A}\boldsymbol{Z})_j}{(\mathbb{A}\boldsymbol{c})_j}, \qquad \mathcal{V}^+(\boldsymbol{Z}) \equiv \min_{j:(\mathbb{A}\boldsymbol{c})_j>0} \frac{b_j - (\mathbb{A}\boldsymbol{Z})_j}{(\mathbb{A}\boldsymbol{c})_j}$$

a

$$\mathcal{V}^0(\boldsymbol{Z}) \equiv \min_{j:(\mathbb{A}\boldsymbol{c})_j=0} b_j - (\mathbb{A}\boldsymbol{Z})_j.$$

Dle Lee a kol. (2016) z lemmatu 4.1 plyne

$$[\boldsymbol{\eta}^\top\boldsymbol{Y} \,|\, \mathbb{A}\boldsymbol{Y} \leq \boldsymbol{b}] \overset{\mathcal{D}}{=} [\boldsymbol{\eta}^\top\boldsymbol{Y} \,|\, \mathcal{V}^-(\boldsymbol{Z}) \leq \boldsymbol{\eta}^\top\boldsymbol{Y} \leq \mathcal{V}^+(\boldsymbol{Z}),\, \mathcal{V}^0(\boldsymbol{Z}) \geq 0]. \tag{4.6}$$

Náhodné veličiny $\mathcal{V}^-(\boldsymbol{Z})$, $\mathcal{V}^0(\boldsymbol{Z})$, $\mathcal{V}^+(\boldsymbol{Z})$ jsou funkcí pouze $\boldsymbol{Z}$. Jelikož jsou náhodné veličiny $\boldsymbol{\eta}^\top\boldsymbol{Y}$ a $\boldsymbol{Z} \equiv (\mathbb{I}_n - \boldsymbol{c}\boldsymbol{\eta}^\top)\boldsymbol{Y}$ nekorelované, čili z normality $\boldsymbol{Y}$ nezávislé, jsou i náhodné veličiny $\mathcal{V}^-(\boldsymbol{Z})$, $\mathcal{V}^0(\boldsymbol{Z})$, $\mathcal{V}^+(\boldsymbol{Z})$ nezávislé s $\boldsymbol{Y}$. Ve vzorci (4.6) lze tedy na $\mathcal{V}^-(\boldsymbol{Z})$, $\mathcal{V}^0(\boldsymbol{Z})$, $\mathcal{V}^+(\boldsymbol{Z})$ nahlížet jako na pevné kvantity a náhodná veličina $[\boldsymbol{\eta}^\top\boldsymbol{Y} \,|\, \mathbb{A}\boldsymbol{Y} \leq \boldsymbol{b}]$ se chová jako náhodná veličina s normálním rozdělením zkráceným na interval $[\mathcal{V}^-(\boldsymbol{Z}),\, \mathcal{V}^+(\boldsymbol{Z})]$. Toto pozorování formalizuje následující věta.

Věta 4.2 (Lee a kol. (2016), Theorem 5.2)**.** *Nechť μ, a, $b \in \mathbb{R}$, $a < b$ a $\sigma^2 \geq 0$. Označme $F^{[a,b]}_{\mu,\sigma^2}(x)$ distribuční funkci náhodné veličiny s rozdělením $\mathcal{N}(\mu, \sigma^2)$, useknutým na interval $[a, b]$, tedy*

$$F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)},$$

kde $\Phi(x)$ značí distribuční funkci normovaného normálního rozdělení $\mathcal{N}(0, 1)$. Nechť $\boldsymbol{Z}$, $\mathcal{V}^-(\boldsymbol{Z})$ a $\mathcal{V}^+(\boldsymbol{Z})$ jsou definovány jako v lemmatu 4.1. Poté má náhodná veličina

$$F^{[\mathcal{V}^-(\boldsymbol{Z}),\mathcal{V}^+(\boldsymbol{Z})]}_{\boldsymbol{\eta}^\top\boldsymbol{\mu},\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta}}(\boldsymbol{\eta}^\top\boldsymbol{Y}|\mathbb{A}\boldsymbol{Y} \leq \boldsymbol{b}) \sim \mathcal{U}(0, 1)$$

rovnoměrné rozdělení na intervalu $(0, 1)$ a

$$[\boldsymbol{\eta}^\top\boldsymbol{Y} \mid \mathbb{A}\boldsymbol{Y} \leq \boldsymbol{b},\, \boldsymbol{Z} = \boldsymbol{z_0}] \sim \mathrm{TN}(\boldsymbol{\eta}^\top\boldsymbol{Y},\, \boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta},\, \mathcal{V}^-(\boldsymbol{Z}_0),\, \mathcal{V}^+(\boldsymbol{z_0})), \tag{4.7}$$

kde $\mathrm{TN}(\boldsymbol{\eta}^\top\boldsymbol{Y},\, \boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta},\, \mathcal{V}^-(\boldsymbol{Z}_0),\, \mathcal{V}^+(\boldsymbol{z_0}))$ značí useknuté normální rozdělení s distribuční funkcí $F^{[\mathcal{V}^-(\boldsymbol{Z}),\mathcal{V}^+(\boldsymbol{Z})]}_{\boldsymbol{\eta}^\top\boldsymbol{\mu},\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta}}$

Poznámka. V Lee a kol. (2016) je zavedena odezva $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2\mathbb{I}_n)$, $\boldsymbol{\mu} \in \mathbb{R}^n$, až do sekce 5.1. V sekci 5.1 je zavedeno $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ a od té doby nám přijde značení rozptylové matice trochu matoucí. V Lee a kol. (2016), Theorem 5.2, je ve vzorci pod (5.9) rozptyl $\sigma^2\|\boldsymbol{\eta}\|_2^2$, který odpovídá $\Sigma = \sigma^2\mathbb{I}_n$ (což nám ještě připadá v pořádku). Dále ale vzorec pod (5.7) obsahuje $\sigma^2\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta}$. Věříme, že jde o přepis, neboť $\mathrm{var}(\boldsymbol{\eta}^\top\boldsymbol{Y}) = \boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta} \stackrel{?}{=} \sigma^2\|\boldsymbol{\eta}\|_2^2$. Zde předpokládáme $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$.

4.2.2 Podmiňování sjednocením polyedrů

Rozdělení $\boldsymbol{\eta}^\top\boldsymbol{Y}$ podmíněné sjednocením náhodných jevů $\bigcup_s\{\mathbb{A}_s\boldsymbol{Y} < \boldsymbol{b}_s\}$ je opět normální, omezené na množinu $\bigcup_s\{\mathbb{A}_s\boldsymbol{Y} < \boldsymbol{b}_s\}$.

Věta 4.3 (Lee a kol. (2016), Theorem 5.3)**.** *Nechť F^G_{μ,σ^2} značí distribuční funkci normálního rozdělení $\mathcal{N}(\mu, \sigma^2)$ omezeného na množinu $G \subseteq \mathbb{R}$. Poté*

$$F^{\bigcup_s[\mathcal{V}^-_s(\boldsymbol{Z}),\mathcal{V}^+_s(\boldsymbol{Z})]}_{\boldsymbol{\eta}^\top\boldsymbol{\mu},\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta}}(\boldsymbol{\eta}^\top\boldsymbol{Y} \mid \bigcup\nolimits_s\{\mathbb{A}_s\boldsymbol{Y} < \boldsymbol{b}_s\}) \sim \mathcal{U}(0, 1),$$

kde $\boldsymbol{Z}$, $\mathcal{V}^-_{\boldsymbol{s}}(\boldsymbol{Z})$ a $\mathcal{V}^+_{\boldsymbol{s}}(\boldsymbol{Z})$ jsou definovány jako v lemmatu 4.1 s $\mathbb{A} = \mathbb{A}_s$ a $\boldsymbol{b} = \boldsymbol{b}_s$.

Využitím Věty 4.3, lemmatu 4.1 a (4.5) získáváme metodu pro konstrukci intervalových odhadů pro regresní koeficient $\beta_j^{\mathcal{M}}$ v modelu zvoleném na základě aktivních složek odhadu metodou lasso, čili $\mathcal{M} = \mathcal{A}(\lambda)$.

Věta 4.4 (Lee a kol. (2016), Theorem 6.1)**.** *Nechť $\boldsymbol{Z}$, $\mathcal{V}^-_{\boldsymbol{s}}(\boldsymbol{Z})$, $\mathcal{V}^+_{\boldsymbol{s}}(\boldsymbol{Z})$ jsou definovány jako ve Větě 4.3 a $\boldsymbol{\eta} = (\mathbb{X}^+_{\mathcal{M}})^\top\boldsymbol{e}_j$. Nechť L a U jsou hodnoty splňující*

$$F^{\bigcup_s[\mathcal{V}^-_s(\boldsymbol{Z}),\mathcal{V}^+_s(\boldsymbol{Z})]}_{L,\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta}}(\boldsymbol{\eta}^\top\boldsymbol{Y}) = 1 - \frac{\alpha}{2}, \qquad F^{\bigcup_s[\mathcal{V}^-_s(\boldsymbol{Z}),\mathcal{V}^+_s(\boldsymbol{Z})]}_{U,\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta}}(\boldsymbol{\eta}^\top\boldsymbol{Y}) = \frac{\alpha}{2}.$$

Poté $[L, U]$ je $(1-\alpha)$-intervalový odhad regresního koeficientu $\beta_j^{\mathcal{M}}$ podmíněně náhodným jevem $\{\mathcal{A}(\lambda) = \mathcal{M}\}$, tedy

$$\mathbb{P}([L, U] \ni \beta_j^{\mathcal{M}} \mid \mathcal{A}(\lambda) = \mathcal{M}) = 1 - \alpha.$$

Jelikož je jev $\{\mathcal{M} = \mathcal{A}(\lambda),\, \boldsymbol{s}(\lambda) = \boldsymbol{s}\}$ podjevem jevu $\{\mathcal{M} = \mathcal{A}(\lambda)\}$, intervalové odhady sestrojené z rozdělení $\boldsymbol{\eta}^\top\boldsymbol{Y} \mid \{\mathcal{M} = \mathcal{A}(\lambda),\, \boldsymbol{s}(\lambda) = \boldsymbol{s}\}$ pomocí Věty 4.2 jsou rovněž validní pro statistickou inferenci po výběru rysů. Dle Lee a kol. (2016) jsou tyto intervalové odhady v porovnání s $\boldsymbol{\eta}^\top\boldsymbol{Y} \mid \{\mathcal{M} = \mathcal{A}(\lambda)\}$ výpočetně nenáročné, ale mohou být znatelně širší.

4.3 Kovarianční test

V této sekci je uveden kovarianční test Lockhart a kol. (2014) signifikance složky vstupující do množiny aktivních složek odhadu metodou lasso. V této sekci budeme předpokládat normální lineární regresní model. Kovarianční test pracuje s cestou odhadu metody lasso $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$, která je vypočítána modifikací algoritmu LARS. Cesta odhadu metodou lasso začíná v uzlu $\lambda_0 = \infty$. Množina aktivních složek odhadu je v tomto uzlu prázdná $\mathcal{A}(\lambda_0) = \emptyset$. S λ klesajícím do nuly se množina aktivních složek odhadu $\mathcal{A}(\lambda)$ mění a v principu převážně zvětšuje (platí $\widehat{\boldsymbol{\beta}}^{Lasso}(0) = \widehat{\boldsymbol{\beta}}^{OLS}$, čili $|\mathcal{A}(0)| = p$). Ke změnám množiny aktivních složek odhadu metodou lasso dochází v uzlech $\lambda_1 > \lambda_2 > \cdots > \lambda_K \geq 0$, $K \in \mathbb{N}$. Kovarianční test testuje signifikanci složky $j \in \mathcal{A}(\lambda_k) \setminus \mathcal{A}(\lambda_{k-1})$, $k \in \{1, \ldots, K\}$, vstupující do množiny aktivních složek odhadu v uzlu λ_k. Hypotézou a alternativou kovariančního testu jsou

$$H_0 : \mathcal{A}(\lambda_{k-1}) \supseteq \mathrm{supp}(\boldsymbol{\beta}^*), \qquad H_1 : j \in \mathrm{supp}(\boldsymbol{\beta}^*).$$

Označme

$$\widetilde{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_{k-1})}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{A}(\lambda_{k-1})|}}{\mathrm{argmin}} \|\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda_{k-1})}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}_{\mathcal{A}(\lambda_{k-1})}\|_1$$

odhad metodou lasso v modelu $\mathcal{A}(\lambda_{k-1}) = \mathcal{A}(\lambda_k) \setminus \{j\}$. Testová statistika kovariančního testu předpokládá známý reziduální rozptyl σ^2 a je definována jako

$$T_k = \frac{\boldsymbol{Y}^\top (\mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_{k+1}) - \mathbb{X}_{\mathcal{A}(\lambda_{k-1})}\widetilde{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_{k-1})}(\lambda_{k+1}))}{\sigma^2}. \tag{4.8}$$

Za předpokladu platnosti nulové hypotézy (tedy podmíněně náhodným jevem $\{\mathcal{A}(\lambda_{k-1}) \supseteq \mathrm{supp}(\boldsymbol{\beta}^*)\}$), jistých předpokladů na vysvětlující proměnné $\boldsymbol{X}$ a předpokladu, že nenulové regresní koeficienty β_j^* jsou velké v absolutní hodnotě, platí

$$T_k \xrightarrow{\mathcal{D}} \mathrm{Exp}(1), \qquad n \to \infty.$$

Poznámka. Testovaná složka j vstupuje do množiny aktivních složek odhadu v uzlu λ_k. Jako intuitivní se tedy jeví porovnávat odhady z testovaných modelů právě v uzlu λ_k. Tento přístup ale nebude fungovat, neboť ze spojitosti odhadu metodou lasso v proměnné λ je $\widehat{\boldsymbol{\beta}}_j^{Lasso}(\lambda_k) = 0$ a z (2.12) plyne

$$\widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_{k-1})}(\lambda_k) = \mathbb{X}^+_{\mathcal{A}(\lambda_{k-1})}(\boldsymbol{Y} - \lambda_{k-1}(\mathbb{X}^\top_{\mathcal{A}(\lambda_{k-1})})^+ \boldsymbol{s}_{\mathcal{A}(\lambda_{k-1})}(\lambda_{k-1})) = \widetilde{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_{k-1})}(\lambda_k),$$

z čehož plyne

$$\mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_k) = \mathbb{X}_{\mathcal{A}(\lambda_{k-1})}\widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_{k-1})}(\lambda_k) = \mathbb{X}_{\mathcal{A}(\lambda_{k-1})}\widetilde{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_{k-1})}(\lambda_k).$$

Hodnota testové statistiky T_k by tedy pro λ_k namísto λ_{k+1} byla rovna nule.

Stále ale není zřejmé, proč odhady $\mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}$ a $\mathbb{X}_{\mathcal{A}(\lambda_{k-1})}\widetilde{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_{k-1})}$ porovnávat právě v následujícím uzlu λ_{k+1}. V Lockhart a kol. (2014) argumentují, že *„j-tý koeficient bude mít v uzlu λ_{k+1} plný efekt na vyrovnanou hodnotu $\mathbb{X}\widehat{\boldsymbol{\beta}}$, těsně před změnou množiny aktivních složek“*. Chceme-li plný efekt, proč netestovat odhady metodou lasso v modelech $\mathcal{A}(\lambda_k)$ a $\mathcal{A}(\lambda_{k-1})$ pro $\lambda = 0$? Poznamenejme proto, že uzly modifikace algoritmu LARS $\lambda_1, \ldots, \lambda_K$ jsou náhodné veličiny a hodnota λ_{k+1} v testové statistice T_k hraje roli času následující události. Asymptotické rozdělení testové statistiky T_k vychází právě z rozdělení náhodných veličin λ_k a λ_{k+1} za platnosti nulové hypotézy. Toto je nejsnazší nahlédnout na speciálním případě ortogonální matice modelu – Lockhart a kol. (2014), sekce 3.1.

V testové statistice T_k v (4.8) je předpokládán známý reziduální rozptyl σ^2. V praxi ovšem bývá reziduální rozptyl σ^2 neznámý a je třeba jej nahradit odhadem. V případě $n > p$ lze použít klasický odhad reziduálního rozptylu, vypočtený pomocí reziduálního součtu čtverců odhadu metodou obyčejných nejmenších čtverců $\widehat{\sigma}^2 = \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}^{OLS}\|_2^2/(n-p)$. Testová statistika

$$F_k = \frac{\boldsymbol{Y}^\top \mathbb{X}\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_{k+1}) - \boldsymbol{Y}^\top \mathbb{X}_{\mathcal{A}(\lambda_{k-1})}\widetilde{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_{k-1})}(\lambda_{k+1})}{\widehat{\sigma}^2}$$

má poté asymptoticky F-rozdělení s dvěma a $(n-p)$ stupni volnosti

$$F_k \xrightarrow{\mathcal{D}} \mathcal{F}(2,\, n-p), \qquad n \to \infty. \tag{4.9}$$

Konvergenci v distribuci (4.9) je snadné nahlédnout, neboť $F_k = T_k/(\widehat{\sigma}^2/\sigma^2)$, přičemž $T_k \xrightarrow{\mathcal{D}} \mathrm{Exp}(1) = \chi^2_2/2$, $n \to \infty$ a $\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$. Statistiky T_k a $\widehat{\sigma}^2$ jsou nezávislé, jelikož:

1. Na statistiku T_k lze nahlížet jako na funkci $\boldsymbol{P}_{\mathbb{X}}\boldsymbol{Y}$ (kde $\boldsymbol{P}_{\mathbb{X}} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top$ je matice ortogonální projekce na lineární prostor $\mathrm{Im}(\mathbb{X})$), neboť modifikace algoritmu LARS pro lasso vydá stejné řešení pro $\boldsymbol{Y}$ i pro $\boldsymbol{P}_{\mathbb{X}}\boldsymbol{Y}$ (zřejmě platí $\mathbb{X}^\top\boldsymbol{P}_{\mathbb{X}}\boldsymbol{Y} = \mathbb{X}^\top\boldsymbol{Y}$, z čehož plyne, že subgradientní podmínka optimality pro metodu lasso (2.3) je stejná pro odezvu $\boldsymbol{Y}$ i pro odezvu $\boldsymbol{P}_{\mathbb{X}}\boldsymbol{Y}$).

2. Statistika $\widehat{\sigma}^2$ je funkcí $(\boldsymbol{I}_n - \boldsymbol{P}_{\mathbb{X}})\boldsymbol{Y}$.

3. Statistiky $\boldsymbol{P}_{\mathbb{X}}\boldsymbol{Y}$ a $(\boldsymbol{I}_n - \boldsymbol{P}_{\mathbb{X}})\boldsymbol{Y}$ jsou nezávislé (Komárek (2021), Theorem 6.2).

Případ $n < p$ již není tak přímočarý. V Lockhart a kol. (2014) uvádějí možnost opět využít odhad $\widehat{\sigma}^2_{\mathcal{M}}$, nyní vypočtený v modelu $\mathcal{M}$, který byl zvolen pomocí křížové validace. Tento přístup ale není podpořen řádnou teorií. Simulace ukazují, že asymptotické rozdělení testové statistiky F_r, kde r je počet parametrů v modelu zvoleném křížovou validací, je blízké rozdělení $F(2,\, n-r)$, ale rozptyl výsledné statistiky může být vyšší a testy mohou být antikonzervativní.

Závěr

V této diplomové práci jsme se zabývali metodou adaptivní lasso (3.1). Odhady vyprodukované metodou adaptivní lasso jsou řídké a dle Věty 3.2 asymptoticky přesně identifikují množinu nulových složek vektoru regresních koeficientů a zároveň jsou $\sqrt{n}$-konzistentními odhady nenulových složek vektoru regresních koeficientů. Díky těmto vlastnostem je metoda adaptivní lasso vhodná do řídkých regresních modelů. V této práci jsme se zabývali pouze lineárním regresním modelem, nicméně většinu uvedených tvrzení lze zobecnit pro zobecněné lineární regresní modely.

V kapitole 1 byl zaveden odhad metodou obyčejných nejmenších čtverců v lineárním regresním modelu. Z výsledků uvedených v sekcích 1.2 a 1.3 vyplynulo, že v případech, kdy počet pozorování n je menší nebo roven, nebo o velmi málo převyšuje počet parametrů modelu p, může být na místě využití vychýlených regularizovaných odhadů. Argumentem k použití regularizovaných odhadů může být i fakt, že i lineární regresní model lze považovat za regularizovaný model, neboť klade restriktivní předpoklady na tvar regresní funkce. V sekci 1.4 jsem se zabývali regularizovanými odhady a uvedli jsme argumenty, proč z ℓ^γ-regularizovaných metod, $\gamma \geq 0$, je do řídkých lineárních regresních modelů nejvhodnější ℓ^1-regularizace, čili metoda lasso. V kapitole 2 byl zaveden odhad metodou lasso. V sekci 2.1 byla odvozena existence, tvar a jednoznačnost odhadu metodou lasso, v sekci 2.2 byla uvedena modifikace algoritmu LARS pro metodu lasso, pomocí které lze vypočítat cestu odhadu metodou lasso. V sekci 2.3 jsme na metodu lasso nahlédli z geometrické perspektivy a z perspektivy Bayesovské statistiky. Sekce 2.4 byla věnována asymptotickým vlastnostem odhadu metodou lasso. Dle Tvrzení 2.6 metoda lasso neidentifikuje konzistentně množinu aktivních složek vektoru regresních koeficientů, což bylo motivací k zavedení metody adaptivní lasso. Kapitola 3 byla věnována odhadu metodou adaptivní lasso. Odhad metodou adaptivní lasso lze získat pomocí metody lasso, aplikujeme-li vhodnou transformaci na matici modelu a vhodnou zpětnou transformaci na výsledný odhadu metodou lasso. Díky tomuto se většina vlastností metody lasso přenáší na metodu adaptivní lasso. Závěrečná kapitola 4 se zabývala statistickou inferencí po výběru rysů v lineárním regresním modelu. Metoda (adaptivní) lasso produkuje řídké odhady a lze ji využít k automatickému výběru rysů v lineárním regresním modelu. V takovém případě jsou ale rysy v lineárním regresním modelu zvoleny na základě pozorovaných dat a klasické metody statistické inference již nejsou zcela validní. Věta 4.3 udává tvar rozdělení odhadu metodou obyčejných nejmenších čtverců, podmíněné modelem zvoleným na základě množiny aktivních složek odhadu metodou (adaptivní) lasso.

Jako hlavní příspěvek této diplomové práce vnímáme shrnutí a zpracování určité problematiky na základě několika různých pramenů. Diskutovanou problematiku doprovázíme různými ilustracemi. Podstatnou část prací jsme strávili na odvození uzavřeného tvaru odhadu metodou lasso v $\mathbb{R}^2$. V pozdější části jsme ale zjistili, že tato vlastnost metody lasso je uvedena již v původním článku Tibshirani (1996). V tomto ohledu si odnášíme i poučení do budoucna.

A. Appendix

A.1 Tvrzení o konvexních funkcích

Lemma A.1.1 (Skládání konvexních funkcí)**.** *Nechť f, $g : \mathbb{R}^n \to \mathbb{R}$ jsou konvexní funkce a $h : \mathbb{R}^p \to \mathbb{R}^n$ je afinní funkce. Poté funkce $f+g : \mathbb{R}^n \to \mathbb{R}$ i $f \circ h : \mathbb{R}^p \to \mathbb{R}$ jsou konvexní.*

Důkaz. Nechť $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^n$, $0 \leq \lambda \leq 1$ a f, $g : \mathbb{R}^n \to \mathbb{R}$ jsou konvexní funkce. Poté

$$\begin{aligned}(f+g)(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) &= f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) + g(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \\ &\leq \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}) + \lambda g(\boldsymbol{x})) + (1-\lambda)g(\boldsymbol{y}) \\ &= \lambda(f+g)(\boldsymbol{x}) + (1-\lambda)(f+g)(\boldsymbol{y})\end{aligned}$$

Nechť $f : \mathbb{R}^n \to \mathbb{R}$ je konvexní funkce, $h : \mathbb{R}^p \to \mathbb{R}^n$ je afinní funkce a $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^p$. Poté

$$\begin{aligned}(f \circ h)(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) &= f(\lambda h(\boldsymbol{x}) + (1-\lambda)h(\boldsymbol{y})) \\ &\leq \lambda f(h(\boldsymbol{x})) + (1-\lambda)f(h(\boldsymbol{y})),\end{aligned}$$

kde jsme v první rovnosti využili toho, že je funkce h afinní a v druhé nerovnosti jsme využili konvexity funkce f.

□

Lemma A.1.2 (Konvexita ℓ^γ norem)**.** *Nechť $\gamma \geq 1$ a $\boldsymbol{x} \in \mathbb{R}^p$. Funkce*

$$\|\boldsymbol{x}\|_\gamma = \left(\sum_{j=1}^{p} |x_j|^\gamma\right)^{1/\gamma}$$

se nazývá ℓ^γ normou vektoru $\boldsymbol{x}$ a je konvexní na $\mathbb{R}^p$.

Důkaz. Normy v $\mathbb{R}^p$ jsou vždy konvexní: pro normu $\|\bullet\| : \mathbb{R}^p \to [0, \infty)$, libovolné $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^p$ a libovolné $0 \leq \lambda \leq 1$ platí

$$\|\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}\| \leq \|\lambda\boldsymbol{x}\| + \|(1-\lambda)\boldsymbol{y}\| = \lambda\|\boldsymbol{x}\| + (1-\lambda)\|\boldsymbol{y}\|,$$

kde jsme využili trojúhelníkové nerovnosti a absolutní homogenity norem. K důkazu tvrzení nám tedy stačí dokázat, že pro $\gamma \geq 1$ je $\|\bullet\|_\gamma$ norma.

Zobrazení $\|\bullet\|_\gamma$ je pozitivně homogenní:

$$\|\alpha\boldsymbol{x}\|_\gamma = \left(\sum_{j=1}^{p} |\alpha x_j|^\gamma\right)^{1/\gamma} = \left(\sum_{j=1}^{p} |\alpha|^\gamma |x_j|^\gamma\right)^{1/\gamma} = |\alpha| \left(\sum_{j=1}^{p} |x_j|^\gamma\right)^{1/\gamma} = |\alpha|\|\boldsymbol{x}\|_\gamma.$$

Subaditivita zobrazení $\|\bullet\|_\gamma$, čili

$$\|\boldsymbol{x} + \boldsymbol{y}\|_\gamma \leq \|\boldsymbol{x}\|_\gamma + \|\boldsymbol{y}\|_\gamma$$

je důsledkem Minkowského nerovnosti (Hardy a kol. (1952), Theorem 25).

□

A.2 Algoritmus LARS for the lasso path

Algorithm 1 Algoritmus *Lars for the lasso path*

▷ Inicializace proměnných
$k \leftarrow 0$ ▷ Počítadlo iterací
$\lambda_0 \leftarrow \infty$ ▷ Počáteční uzel
$\boldsymbol{s}(\lambda_0) \leftarrow \mathbf{0}_p$ ▷ Vektor znamének současného odhadu
$\mathcal{A}(\lambda_0) \leftarrow \emptyset$ ▷ Množina aktivních složek současného odhadu
while $\lambda_k > 0$ **do**
  $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda_k) \leftarrow \mathbb{X}^{+}_{\mathcal{A}(\lambda_k)}(\boldsymbol{Y} - \lambda_k(\mathbb{X}^{\top}_{\mathcal{A}(\lambda_k)})^{+}\boldsymbol{s}(\lambda_k))$ ▷ Současný odhad
  for $j \notin \mathcal{A}(\lambda_k)$ **do**
    $t_j^{join+} \leftarrow [\boldsymbol{X}_j^{\top}(\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda_k)}\mathbb{X}^{+}_{\mathcal{A}(\lambda_k)}\boldsymbol{Y})]/[1 - \boldsymbol{X}_j^{\top}(\mathbb{X}^{\top}_{\mathcal{A}(\lambda_k)})^{+}\boldsymbol{s}(\lambda_k)]$
    if $0 \leq t_j^{join+} < \lambda_k$ **then**
      $t_j^{join} \leftarrow t_j^{join+}$
    else
      $t_j^{join} \leftarrow [\boldsymbol{X}_j^{\top}(\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda_k)}\mathbb{X}^{+}_{\mathcal{A}(\lambda_k)}\boldsymbol{Y})]/[-1 - \boldsymbol{X}_j^{\top}(\mathbb{X}^{\top}_{\mathcal{A}(\lambda_k)})^{+}\boldsymbol{s}(\lambda_k)]$
    end if
  end for
  $\lambda_{k+1}^{join} = \max_{j \notin \mathcal{A}(\lambda_k)} t_j^{join}$
  for $j \in \mathcal{A}(\lambda_k)$ **do** ▷ Výpočet prvního času odchodu
    $t_j^{leave} \leftarrow [\mathbb{X}^{+}_{\mathcal{A}(\lambda_k)}\boldsymbol{Y}]_j/[(\mathbb{X}^{\top}_{\mathcal{A}(\lambda_k)}\mathbb{X}_{\mathcal{A}(\lambda_k)})^{+}\boldsymbol{s}(\lambda_k)]_j$
  end for
  $\lambda_k^{leave} \leftarrow \max_{j \in \mathcal{A}(\lambda_k)} t_j^{leave}$
  if $\lambda_k^{join} > \lambda_k^{leave}$ **then** ▷ Výpočet nového uzlu a aktualizace znamének
    $\lambda_{k+1} \leftarrow \lambda_k^{join}$
    $j_{k+1}^{join} \leftarrow \operatorname{argmax}_{j \notin \mathcal{A}(\lambda_k)} t_j^{join}$
    $s_{k+1}^{join} \leftarrow \operatorname{sign}(\boldsymbol{X}^{\top}_{j_{k+1}^{join}}(\boldsymbol{Y} - \mathbb{X}_{\mathcal{A}(\lambda_k)}\widehat{\boldsymbol{\beta}}^{Lasso}_{\mathcal{A}(\lambda_k)}(\lambda)))$
    $\boldsymbol{s}(\lambda_{k+1}) \leftarrow \boldsymbol{s}(\lambda_k) + s_{k+1}^{join}\boldsymbol{e}_{j_{k+1}^{join}}$
    $\mathcal{A}(\lambda_{k+1}) \leftarrow \mathcal{A}(\lambda_k) \cup \{j_{k+1}^{join}\}$
  else
    $\lambda_{k+1} \leftarrow \lambda_k^{leave}$
    $j_{k+1}^{leave} \leftarrow \operatorname{argmax}_{j \in \mathcal{A}(\lambda_k)} t_j^{leave}$
    $s_{k+1}^{leave} \leftarrow s_{j_{k+1}^{leave}}(\lambda_k)$
    $\boldsymbol{s}(\lambda_{k+1}) \leftarrow \boldsymbol{s}(\lambda_k) - s_{k+1}^{leave}\boldsymbol{e}_{j_{k+1}^{leave}}$
    $\mathcal{A}(\lambda_{k+1}) \leftarrow \mathcal{A}(\lambda_k) \setminus \{j_{k+1}^{leave}\}$
  end if
end while

Seznam použité literatury

Andrews, D. F. a Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, **36** (1), 99–102. URL https://www.jstor.org/stable/2984774.

Austin, P. C. a Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of clinical epidemiology*, **68**(6), 627–636. URL https://doi.org/10.1016/j.jclinepi.2014.12.014.

Barto, L. a Tůma, J. (2019). *Lineární Algebra*. URL https://www.karlin.mff.cuni.cz/~barto/LinAlg/skripta_la6.pdf.

Berk, R., Brown, L., Buja, A., Zhang, K. a Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, pages 802–837. URL https://doi.org/10.48550/arXiv.1306.1059.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37**(4), 373–384.

Efron, B., Hastie, T., Johnstone, I. a Tibshirani, R. (2004). Least angle regression. URL https://doi.org/10.48550/arXiv.math/0406456.

Fan, J. a Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96** (456), 1348–1360. URL https://doi.org/10.1198/016214501753382273.

Fu, W. a Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics*, **28**(5), 1356–1378. URL https://www.jstor.org/stable/2674097.

Groves, T. a Rothenberg, T. (1969). A note on the expected value of an inverse matrix. *Biometrika*, **56**(3), 690–691. URL https://doi.org/10.1093/biomet/56.3.690.

Hardy, G. H., Littlewood, J. E. a Pólya, G. (1952). *Inequalities*. Cambridge university press.

Hastie, T., Tibshirani, R., Friedman, J. H. a Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer. URL https://hastie.su.domains/Papers/ESLII.pdf.

Komárek, A. (2021). *Course Notes – NMSA407 Linear Regression*. URL https://www.karlin.mff.cuni.cz/~kulich/vyuka/linreg/doc/2021-NMSA407-notes.pdf.

Kulich, M. (2022). *Extended Course Notes – NMST432 Advanced Regression Models*. URL https://www.karlin.mff.cuni.cz/~kulich/vyuka/pokreg/doc/advreg_notes_ext_220218.pdf.

Lachout, P. (2020). *Optimization Theory - direct approach*.

Lee, J. D., Sun, D. L., Sun, Y. a Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. URL https://doi.org/10.48550/arXiv.1311.6238.

Leeb, H. a Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, **21**(1), 21–59. URL https://www.jstor.org/stable/3533623.

Leeb, H. a Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of hodges' estimator. *Journal of Econometrics*, **142**(1), 201–211. URL https://doi.org/10.48550/arXiv.0704.1466.

Levine, R. A. a Casella, G. (2001). Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, **10**(3), 422–439. URL https://doi.org/10.1198/106186001317115045.

Lockhart, R., Taylor, J., Tibshirani, R. J. a Tibshirani, R. (2014). A significance test for the lasso. *Annals of statistics*, **42**(2), 413. URL https://doi.org/10.48550/arXiv.1301.7161.

Mairal, J. a Yu, B. (2012). Complexity analysis of the lasso regularization path. URL https://doi.org/10.48550/arXiv.1205.0079.

Osborne, M. R., Presnell, B. a Turlach, B. A. (2000a). On the lasso and its dual. *Journal of Computational and Graphical statistics*, pages 319–337. URL https://www.jstor.org/stable/1390657.

Osborne, M. R., Presnell, B. a Turlach, B. A. (2000b). A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, **20**(3), 389–403. URL https://doi.org/10.1093/imanum/20.3.389.

Park, T. a Casella, G. (2008). The bayesian lasso. *Journal of the american statistical association*, **103**(482), 681–686. URL https://doi.org/10.1198/016214508000000337.

Raskutti, G., Wainwright, M. J. a Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, **57**(10), 6976–6994. URL https://doi.org/10.48550/arXiv.0910.2042.

Robert, C. P. a Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Ročková, V. a George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, **113**(521), 431–444. URL https://doi.org/10.1080/01621459.2016.1260469.

Stoica, P. a Ottersten, B. (1996). The evil of superefficiency. *Signal Processing*, **55**(1), 133–136. URL https://doi.org/10.1016/S0165-1684(96)00159-4.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**(1), 267–288. URL https://www.jstor.org/stable/2346178.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, **7**(none), 1456 – 1490. doi: 10.1214/13-EJS815. URL https://doi.org/10.48550/arXiv.1206.0313.

Wasserman, L. 36-708 statistical methods for machine learning, minimax theory. URL https://www.stat.cmu.edu/~larry/=sml/minimax.pdf.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**(476), 1418–1429. URL https://doi.org/10.1198/016214506000000735.

Seznam obrázků

1.1 Testovací a trénovací chyba 10
1.2 Příklady regularizovaných odhadů 14
1.3 Střední čtvercová chyba odhadu hřebenovou regresí 15
1.4 Škálovaná střední čtvercová chyba regularizovaných odhadů 17

2.1 Cesta odhadu metodou lasso v $\mathbb{R}^2$, normované vysvětlující proměnné 24
2.2 Patologický případ cesty odhadu metodou lasso v $\mathbb{R}^2$ 25
2.3 Patologický případ cesty vyrovnaných hodnot odhadu metodou lasso v $\mathbb{R}^2$ 25
2.4 Příklady aposteriorních hustot pro lasso 27
2.5 Porovnání regularizačních parametrů pro lasso 28
2.6 Bimodální sdružená aposteriorní hustota pro $\boldsymbol{\beta}$, σ^2 29

3.1 Příklady řídkých adaptivních odhadů 33
3.2 Cesta odhadu metodou lasso a transformace na cestu odhadu metodou adaptivní lasso 34