SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
Fakulta elektrotechniky a informatiky

Jozef Jakubík

Autoreferát dizertačnej práce

Metódy analýzy vysoko dimenzionálnych dát pomocou lineárnych zmiešaných modelov

na získanie akademickej hodnosti doktor (philosophiae doctor, PhD.)

Ústav merania
Slovenskej akadémie vied

Študijný program: meracia technika
Študijný odbor: 5.2.54. meracia technika

Bratislava 2018

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Ústave merania, Slovenskej akadémie vied.

Doktorand: Jozef Jakubík
Ústav merania SAV
Dúbravská cesta 9
841 04 Bratislava

Školiteľ: Doc. RNDr. Viktor Witkovský, CSc.
Ústav merania SAV
Dúbravská cesta 9
841 04 Bratislava

Oponenti: Doc. RNDr. Katarína Kozlíková, CSc.
Lekárska fakulta UK
Špitálska 24
813 72 Bratislava

prof. RNDr. Gejza Wimmer, DrSc.
Matematický ústav SAV
Štefánikova 49
814 73 BRATISLAVA

Autoreferát bol rozposlaný dňa: . . . . . .
Obhajoba dizertačnej práce sa koná dňa 28. 8. 2018 o 11:00 hod. v zasadačke Ústavu merania SAV, Dúbravská cesta 9, 841 04 Bratislava.

. . . . . .
prof. Dr. Ing. Miloš Oravec
Dekan FEI STU v Bratislave

Obsah

Úvod

Metódy štatistickej analýzy dát (meraní) pomocou lineárnych resp. nelineárnych regresných modelov patria medzi najčastejšie používané metódy modelovania v mnohých oblastiach vedeckého experimentálneho výskumu. Moderné metódy v oblasti technického ako aj biomedicínskeho výskumu umožňujú charakterizáciu každého subjektu v danom experimente pomocou veľkého množstva potenciálnych vysvetľujúcich premenných - regresorov respektíve prediktorov uvažovaného matematického modelu. Príkladom takýchto dát môžu byť napríklad spektrometrické merania subjektov experimentu pomocou spektrometra/chromatografu, EEG merania, alebo výsledky analýzy genetických DNA-čipových experimentov (DNA Microarray Experiments). V takýchto situáciách je počet subjektov v experimente omnoho menší ako počet vysvetľujúcich premenných pre každý subjekt a priama aplikácia štandardných postupov štatistickej inferencie nie je možná. V posledných rokoch sa dosiahol významný pokrok v oblasti rozvoja teórie, metód a algoritmov pre odhadovanie parametrov vo vysoko dimenzionálnych lineárnych regresných modeloch. Klasický predpoklad o stochastickej nezávislosti pozorovaní býva často nereálny a jednou z možností ako modelovať komplikovanú kovariančnú štruktúru medzi skupinami subjektov v experimente je využitie lineárnych zmiešaných modelov.

Práca sa zoberá problematikou výberu vhodných regresorov ovplyvňujúcich pozorovania vo vysoko dimenzionálnych dátach a poskytuje prehľad prác v oblasti vysoko dimenzionálnych lineárnych modelov a lineárnych zmiešaných modelov. Analýza vysoko dimenzionálnych dát je motivovaná napríklad genetikou, kde GWAS - genome-wide association štúdie, sú zamerané na analýzu celého genómu, ktorý môže obsahovať aj viac ako milión regresorov, a jeho vplyv na pozorovania [Zhou et al., 2013, Tan et al., 2018]. Z hľadiska interpretovania výsledkov je nevhodné uvažovať, že celý genóm ovplyvňuje pozorovaný znak a preto je často potrebné vybrať malú skupinu regresorov, ktoré majú vplyv na pozorovania. Ďalšie uplatnenie nachádzame v oblasti spektrometrického merania subjektov [Kushch et al., 2008, Schwarz et al., 2009]. Výsledkom spektrometrického merania môže byť spektrum zložiek vydychovaného vzduchu pacienta, ktoré môže obsahovať tisíce prvkov, čo je výrazne viac ako počet pacientov, ktorých máme k dispozícii. Podobne formulovateľné problémy sa dajú nájsť napríklad v meteorológii [Pampuri et al., 2011], ekonómii [Bai and Ng, 2008] a mnohých ďalších oblastiach.

K problematike vysoko dimenzionálnych dát existujú rôzne prístupy. Niekoho nezaujíma výber regresorov, ale má záujem len o predikciu alebo odhad parametrov modelu. V predloženej práci sa venujeme najmä problematike výberu regresorov, ktorých efekt môžeme následne odhadnúť pomocou bežne zaužívaných odhadov, najmä z toho dôvodu, že v praktických prípadoch, keď sa dimenzia problému blíži až k 10^6, je často veľký problém nájsť kompromis medzi veľkosťou množiny vybraných regresorov, ktoré vplývajú na pozorovania a presnosťou odhadu parametrov.

Prvá kapitola sa venuje predstaveniu klasického lineárneho zmiešaného modelu a základných prístupov na odhadovanie parametrov, z ktorých sa neskôr vychádza pri odvodení metód používaných vo vysoko dimenzionálnom prípade. V oblasti lineárnych regresných modelov sa problematika vysoko dimenzionálnych dát aktívne študuje už nejakú dobu, prehľad dosiahnutých výsledkov a navrhnutých prístupov v tejto oblasti je zhrnutý v druhej kapitole. Tretia ka-

pitola nadväzuje na predchádzajúce dve a uvádza prehľad prístupov, ktoré boli navrhnuté na analýzu vysoko dimenzionálnych lineárnych zmiešaných modelov. Doposiaľ navrhnuté metódy sú ale výpočtovo náročné a je problém ich využiť v prípadoch s veľkou dimenziou.

V nasledujúcej kapitole preto prinášame návrh postupu na výber regresorov, ktorý sa dá uplatniť v prípade vysokých dimenzií. Prezentovaný je postupný vývin metódy od najjednoduchšieho prístupu až po komplexnejšie prístupy. V nasledujúcej teoretickej časti následne odvodíme vlastnosť konzistencie ktorá zabezpečí, že s rastúcim počtom pozorovaní navrhnuté metódy určite nájdu požadovanú množinu regresorov.

V posledných kapitolách odprezentujeme výsledky konkrétnych príkladov a simulačných štúdií, v ktorých porovnáme navrhnuté metódy s už existujúcimi metódami a zároveň ukážeme správanie sa metód v rôznych situáciach.

Tézy dizertačnej práce

- Preštudovať súčasné metódy na výber regresorov vo vysoko dimenzionálnych lineárnych zmiešaných modeloch (LMM), navrhnúť novú efektívnu metódu a teoreticky podložiť navrhnutú metódu. Súčasná literatúra venujúca sa problematike vysoko dimenzionálnych dát zahŕňa okrem algoritmov aj teoretické základy z ktorých vychádzajú. Často sa vyskytujú tvrdenia, ktoré sú odvodené za predpokladu, že skutočnosť poznáme, ktoré dokladajú, že daná metóda nájde pri vhodných parametroch presnú množinu regresorov s veľkou pravdepodobnosťou. Cieľom dizertačnej práce je preskúmať a porovnať vlastnosti týchto metód aj keď tieto parametre nepoznáme.

- Problém súčasného softvéru na analýzu vysoko dimenzionálnych dát je najmä vo výpočtovej náročnosti, ktorá neumožňuje jeho praktické využitie v prípadoch, keď sa dimenzia problému blíži k 10^4. Preto je cieľom dizertačnej práce navrhnutú metódu implementovať tak aby sa dala použiť na problémy s dimenziou 10^5.

- Dôkladne analyzovať a porovnať navrhnutú metódu s metódami, ktoré existujú v súčasnosti. Na základe simulačných štúdií na náhodne vygenerovaných dátach ako aj dátach odvodených z reálnych dát porovnať vytvorenú metódu s metódami, ktoré existujú na vysoko dimenzionálnych problémoch s nižšou dimenziou. V prípade problémov s vyššou dimenziou porovnať výsledky so známou skutočnosťou. Zároveň analyzovať správanie sa metódy pri zmene parametrov pôvodného modelu.

1 Lineárny zmiešaný model (LMM)

Lineárny zmiešaný model (LMM - linear mixed model) [Eisenhart, 1947] je úzko spojený s modelmi analýzy rozptylu. LMM umožňujú špecifikovať štruktúru kovariančnej matice daného modelu. Táto informácia môže byť dôležitá pri ďalšom štatistickom vyhodnotení dát. LMM majú široké uplatnenie napríklad v technických vedách, biomedicíne, behaviorálnych výskumoch.

LMM v najvšeobecnejšej forme [Searle et al., 1992], [West et al., 2006] môžeme zapísať:

$$\mathrm{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathrm{u} + \varepsilon, \tag{1-1}$$

kde

Y je $\mathrm{n} \times 1$ vektor pozorovaní.

$\mathbf{X}$ je $\mathrm{n} \times \mathrm{p}$ matica regresorov.

$\boldsymbol{\beta}$ je $\mathrm{p} \times 1$ vektor neznámych pevných efektov.

$\mathbf{Z}$ je $\mathrm{n} \times \mathrm{q}$ matica prediktorov.

u je $\mathrm{q} \times 1$ vektor náhodných efektov z rozdelenia $\mathcal{N}(0,\mathbf{D})$, s neznámou variačno-kovariačnou maticou $\mathbf{D}$.

ε je $\mathrm{n} \times 1$ realizovaný vektor chýb z rozdelenia $\mathcal{N}(0,\mathbf{R})$, kde matica $\mathbf{R}$ je taktiež neznáma. ε je nezávislý od u.

Pod efektom v práci chápeme koeficient prislúchajúci k danému regresoru, resp. prediktoru v príslušnom odhadovanom vektore.

Často sa predpokladá, že variančno-kovariančná matica $\mathbf{D}$ nie je hustá, ale má blokovo diagonálnu až diagonálnu štruktúru, pričom veľkosti jednotlivých blokov sú známe. Rovnako sa často predpokladá, že matica $\mathbf{R} = \sigma^2\mathbf{I}$ a nezávislosť vektorov $\mathbf{u}$ a $\boldsymbol{\varepsilon}$. V nasledujúcom texte budeme uvádzať náhodné premenné rovnakým fontom ako pozorovania.

2 Výber regresorov vo vysoko dimenzionálnom lineárnom modeli

Vysoko dimenzionálnym dátam [Friedman et al., 2009] sa v súčasnosti venuje veľká pozornosť najmä v bioinformatike, informačných technológiách alebo astronómii. „Vysoká dimenzionalita“ sa vzťahuje na fakt, že počet neznámych parametrov, ktoré chceme odhadnúť, niekoľkonásobne prevyšuje počet pozorovaní ktoré máme k dispozícii. Ako príklad uvedieme biologické dáta z práce [Atwell et al., 2010], v ktorej študujú závislosť fenotypu, vonkajších prejavov, od genotypu, dedičných faktorov, u rastliny *Arabidopsis thaliana*. Skúmané dáta pozostávajú zo SNP-ov (Single-nucleotide polymorphism), čo sú miesta v reťazci DNA, na ktorých pozorujeme v populácii rôzne nukleotidy. Napríklad, ak máme v populácii aspoň jedného jedinca,

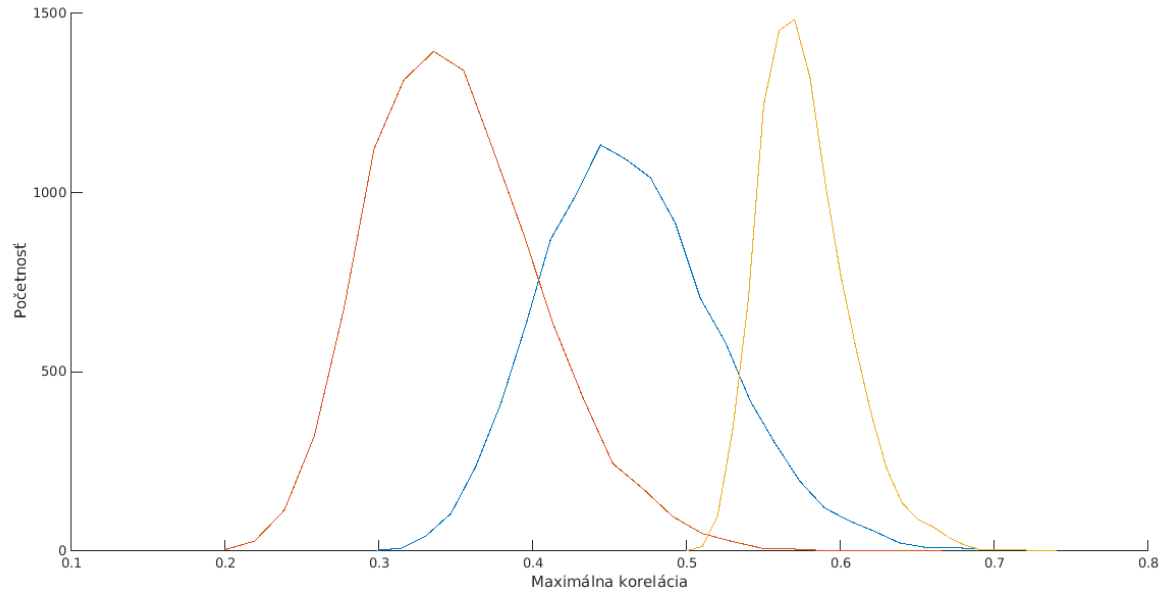
ktorý má v DNA namiesto väčšinového úseku ...AAGCCTA... úsek ...AAGCTTA..., tak podčiarknuté mieste nazveme SNP. Počet SNP u rastliny *Arabidopsis thaliana* dosahuje 250000, pričom počet skúmaných rastlín je asi 200. Matica SNP-ov, ktorá zachytáva aký nukleotid sa u daného jedinca nachádza na danom SNP-e, sa transformuje na číselné hodnoty a normalizuje. Predpokladajme, že pozorujeme znak ktorý je závislý na genotype. Uvažujme, že sa táto závislosť dá modelovať pomocou lineárnej regresie:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\mathbf{Y}$ je n-rozmerný vektor pozorovaní, $\mathbf{X}$ je $n \times p$ matica SNP-ov, $\boldsymbol{\beta}$ je p-rozmerný vektor neznámych efektov prislúchajúcich k jednotlivým SNP-om a $\boldsymbol{\varepsilon}$ je vektor chýb s rozdelením $\mathcal{N}(0, \sigma^2\mathbf{I}_n)$. Vidíme, že počet neznámych regresorov $p = 250000$ výrazne prevyšuje počet pozorovaní $n = 200$. Budeme predpokladať, že nie všetky SNP ovplyvňujú nami pozorovaný znak. Našou snahou je nájsť množinu SNP-ov $S \subset \{1, \ldots, p\}$, ktorá by mala byť rovnaká ako množina $S^0 \subset \{1, \ldots, p\}$, čo je neznáma skutočná množina SNP-ov, ktoré ovplyvňujú nami pozorovaný znak. Počet týchto SNP nech je $s^0 = |S^0|$.

Takto zadaný problém má technické ako aj teoretické obmedzenia [Fan and Lv, 2010]. Nie je možné použiť bežné metódy, napríklad metódu najmenších štvorcov, lebo by viedla k nejednoznačným odhadom. Problémom je aj použitie metód výberu regresorov, ktoré sú navrhnuté pre dáta s nízkou dimenziou, lebo s rastúcou dimenziou rýchlo rastie výpočtová zložitosť. Často to súvisí najmä s tým, že dané metódy optimalizujú nekonvexnú účelovú funkciu. Existujú ale aj metódy, ktoré dokážu riešiť problémy s dimenziou viac ako 10^6. Komplikovanejšie metódy majú často lepší teoretický základ a v nižších dimenziách dávajú aj lepšie praktické výsledky. V súčasnosti sa už problémy často delia podľa dimenzie aj na „vysoko dimenzionálne“ a „ultra vysoko dimenzionálne“.

Kým technické obmedzenia sa dajú obísť zlepšením algoritmov a výpočtovej techniky, teoretické obmedzenia sú oveľa zväzujúcejšie. Teoretické obmedzenia súvisia najmä s prirodzenou tendenciou metód uprednostňovať pri výbere regresory, ktoré sú viac korelované s pozorovaným znakom. Podobný problém nastáva pri testovaní veľkého počtu hypotéz, ktorý je známy ako *multiple comparisons* alebo *multiple testing problem* (problém viacnásobného testovania). V prípade výberu regresorov nastáva problém, že s veľkým počtom regresorov sa zvyšuje šanca, že niektorý regresor, ktorý nemá vplyv na pozorovania, bude mať príliš veľkú koreláciu s pozorovaniami. Napríklad predpokladajme úlohu v ktorej máme $n = 50$ pozorovaní. Pozorovania závisia od dvadsiatich regresorov ($s^0 = 20$), od každého rovnako. My ale pozorujeme $p_1 = 100000$ regresorov respektíve $p_2 = 100$ regresorov. Desaťtisíc krát sme generovali zmienenú úlohu a zaznamenali si maximálnu koreláciu pozorovaní s regresormi od ktorých pozorovania závisia a so všetkými pozorovaniami $\max_i(\rho_{\mathbf{Y},\mathbf{X}_{(:,i)}})$, kde i prechádza cez množinu S^0 pre regresory od ktorých pozorovania závisia a množiny $\{1, \ldots, p_1\}$ respektíve $\{1, \ldots, p_2\}$ všetkých regresorov. Regresory sú generované z normálneho rozdelenia a sú všetky navzájom nezávislé. Z obrázku 2 -- 1 je zrejmé, že pri veľkom počte regresorov často nastáva situácia, kde regresory s najväčšou koreláciou neovplyvňujú pozorovania. Prirodzene sa potom v prípade väčšieho počtu regresorov budú metódy, ktoré na výber použijeme, častejšie mýliť, lebo prirodzene uprednostňujú regresory viac korelované s pozorovaniami. Práca [Bühlmann and



Obr. 2 – 1 Porovnanie histogramu maximálnej korelácie pozorovaní s dvadsiatimi regresormi, od ktorých pozorovania závisia (modrá), s histogramami maximálnej korelácie pozorovaní so sto (červená) a stotisíc (žltá) regresormi od ktorých sú pozorovania nezávislé.

van de Geer, 2011] prichádza s podmienkou

$$\log(p) \cdot s^0 \ll n,$$

pri ktorej má väčšinou zmysel uvažovať o výbere regresorov vo vysoko dimenzionálnych dátach.

2.1 Často používané metódy pri výbere regresorov

V klasickej lineárnej regresii $(p < n)$ môžeme testovať významnosť jednotlivých regresorov a na základe toho potom model zjednodušiť a pomocou vhodného kritéria rozhodnúť, či je nový model vhodnejší ako pôvodný. V modeloch s malou dimenziou nie je problém zopakovať takýto postup aj viackrát. Problém nastáva s rastúcou dimenziou, keďže počet všetkých potenciálnych modelov je 2^p. V nasledujúcej časti uvedieme prehľad najpoužívanejších prístupov v oblasti výberu regresorov vo vysoko dimenzionálnych modeloch. Mnoho uvedených metód sa často používa nielen na výber regresorov, ale aj priamo na odhad efektov týchto regresorov. Takýto postup má význam v prípade nižšej dimenzie dát, ale s rastúcou dimenziou nastáva problém s tým, že ak požadujeme množinu regresorov s malou mohutnosťou, tak je odhad efektov silno vychýlený. V prípade metód, ktoré odhadujú efekty potom za vybrané regresory považujeme tie, ktoré majú v danom odhade nenulový efekt (v absolútnej hodnote väčší ako nami zvolená konštanta).

2.1.1 Penalizované odhady

V súčasnosti veľmi obľúbené metódy na výber regresorov, ktoré zároveň umožňuje aj odhad efektov týchto regresorov. Tieto metódy spočívajú v tom, že penalizujú klasický odhad najmenších štvorcov

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda f(\boldsymbol{\beta}),$$

kde $f(\boldsymbol{\beta})$ je penalizačná funkcia a λ je parameter, ktorý prikladá váhu k danej penalizácii.

Asi najpopulárnejšia je ℓ_1 penalizácia, ktorá vedie k metóde *LASSO* - Least Absolute Shrinkage and Selection Operator (laso metóda) [Tibshirani, 1996].

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2-1}$$

Obľúbená je najmä kvôli kvadratickej (konkávnej) účelovej funkcii, vďaka čomu sa dá efektívne uplatniť aj na dáta s veľmi vysokou dimenziou. Funkcia `lasso( )` v MATLAB-e dokáže odhadnúť efekty v úlohách, v ktorých počet regresorov dosahuje až 10^6.

Ďalšia obľúbená metóda je ℓ_2 penalizovaný odhad, v literatúre tiež známy ako Tikhonova regularizácia, alebo *ridge* (hrebeňová) regresia [Waldron et al., 2011].

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2, \tag{2-2}$$

Na výber regresorov nie je až tak populárna, lebo jej výsledkom nie je riedky odhad efektov na základe ktorého by sme vybrali regresory.

3 Výber regresorov vo vysoko dimenzionálnych lineárnych zmiešaných modeloch

Prirodzeným zovšeobecnením vysoko dimenzionálnych lineárnych modelov z predchádzajúcej kapitoly je (najmä v oblasti genetických dát) vysoko dimenzionálny lineárny zmiešaný model. Predpokladajme klasický LMM, ako bol definovaný v 1-1, len s tým rozdielom, že dimenzia p matice $\mathbf{X}$ je väčšia ako počet pozorovaní n ($n < p$). My budeme uvažovať len prípady, keď sa „vysoká dimenzia“ vzťahuje len na maticu $\mathbf{X}$ prislúchajúcu k pevným efektom. $\mathbf{X}$ zachytáva napríklad jednotlivé gény, z ktorých chceme vybrať tie, ktoré majú vplyv na pozorovania. Matica $\mathbf{Z}$ často zachytáva príbuzenské vzťahy alebo iné vonkajšie faktory.

Na takto zadaný problém sa dá uplatniť viacero upravených metód spomenutých v predchádzajúcej kapitole. Asi najpopulárnejšia je opäť penalizácia pomocou ℓ_1 normy. V článku [Schelldorfer et al., 2011] autori navrhli, za predpokladov, že $\mathbf{R} = \sigma^2\mathbf{I}$, vektory $\boldsymbol{\varepsilon}$ a $\mathbf{u}$ sú nezávislé a $\mathbf{D}$ je pozitívne definitná a závisí od parametrov θ (napríklad cez Choleského rozklad), ktorých je menej ako pozorovaní n, odhad odvodený z ML odhadu vektora $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} = (\mathbf{Z}\mathbf{D}\mathbf{Z}^\mathsf{T} + \mathrm{R}))$, ktorý penalizujú ℓ_1 normou:

$$(\hat{\boldsymbol{\beta}}, \hat{\theta}, \hat{\sigma}^2) = \arg\min_{\boldsymbol{\beta},\theta,\sigma^2>0,D\succ 0} \left[\frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\right]. \tag{3-1}$$

V takomto prípade je účelová funkcia vo všeobecnosti nekonvexná. Výhody konvexnej účelovej funkcie spočívajú v menšej výpočtovej zložitosti, ale najmä v garancii globálneho minima. Pre takýto odhad bolo v článku odvodených viacero teoretických výsledkov, spomedzi ktorých najpodstatnejší je (pri technických predpokladoch uvedených v článku), že pre vhodne zvolené λ s veľkou pravdepodobnosťou platí

$$S^0 \subset \hat{S}_\lambda = \left\{1 \leq k \leq p : \hat{\beta}_{k,\lambda} \neq 0\right\}$$

a množina $\hat{S}_\lambda$ nemá príliš veľkú mohutnosť v porovnaní s mohutnosťou množiny S^0 .

Ďalej autori rozšírili svoje teoretické výsledky aj na adaptívny odhad

$$(\hat{\boldsymbol{\beta}}, \hat{\theta}, \hat{\sigma}^2) = \underset{\boldsymbol{\beta},\theta,\sigma^2>0,D\succ 0}{\arg\min} \left[\frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{w}^T\boldsymbol{\beta}\|_1\right],$$

kde $\boldsymbol{w}$ je vektor váh, podobne ako v adaptívnom LASSO odhade z predchádzajúcej kapitoly. Autori taktiež navrhli a implementovali algoritmus v jazyku R, kde je dostupný v balíku `lmm–lasso`.

V článku [Rohart et al., 2014] autori pri podobných predpokladoch navrhujú metódu vychádzajúcu z ML odhadu parametrov združného rozdelenia vektora $\left[\mathbf{Y}^T, \mathbf{u}^T\right]^T$

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{D}}, \hat{\sigma}^2) = \underset{\boldsymbol{\beta},\mathbf{D},\sigma^2>0}{\arg\min} \left[\frac{1}{2}\log|\mathbf{R}| + \frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^T\mathbf{R}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) + \right. \tag{3-2}$$

$$\left. \frac{1}{2}\log|\mathbf{D}| + \frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u} + \lambda\|\boldsymbol{\beta}\|_1\right]. \tag{3-3}$$

V dizertačnej práci [Lippert, 2013], v ktorej sa autor zaoberá hlavne rýchlym odhadovaním veľkých lineárnych zmiešaných modelov, sa autor venuje aj výberu regresorov. Autor pracuje s lineárnym zmiešaným modelom, ktorý má dva variančné komponenty, označme ich σ_D^2, pričom $\mathbf{D} = \sigma_D^2\mathbf{K}$, kde $\mathbf{K}$ je známa symetrická matica, a σ^2, pričom $\mathbf{R} = \sigma^2\mathbf{I}$. Postupuje ako v Bayesovskej štatistike a uvažuje, že odhad $\boldsymbol{\beta}$ na základe daných dát je len jedným možným prejavom posteriórnej distribučnej funkcie s riedkym priorom. Uvažuje symbolicky zapísanú posteriórnu funkciu

$$\mathcal{N}(\mathbf{Y}|\mathbf{X}\boldsymbol{\beta}, \sigma_D^2\mathbf{K} + \sigma^2\mathbf{I}) \prod_{j=1}^{p} \exp(-\lambda|\beta_j|)$$

Najskôr pomocou ML odhadu odhadne $\hat{\gamma}$, kde $\gamma = \sigma_D^2/\sigma^2$, pri „nulovom modeli“, t.j. modeli, v ktorom sa ignorujú všetky regresory, a následne transformuje dáta

$$\tilde{\mathbf{X}} = (\hat{\gamma}\boldsymbol{\Lambda} + \mathbf{I})^{-\frac{1}{2}}\mathbf{U}^T\mathbf{X}$$
$$\tilde{\mathbf{Y}} = (\hat{\gamma}\boldsymbol{\Lambda} + \mathbf{I})^{-\frac{1}{2}}\mathbf{U}^T\mathbf{Y},$$

kde $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ je spektrálny rozklad. Táto transformácia eliminuje koreláciu medzi pozorovaniami a transformuje kovariančnú maticu

$$\mathcal{N}(\mathbf{Y}|\mathbf{X}\boldsymbol{\beta}, \gamma\boldsymbol{\Lambda} + \mathbf{I}) \prod_{j=1}^{p} \exp(-\lambda|\beta_j|)$$

a hľadanie hodnoty, ktorá maximalizuje danú posteriórnu funkciu je ekvivalentné LASSO odhadu.

$$\min_{\beta} \frac{1}{\sigma^2}\|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Ďalšie metódy, ktoré boli navrhnuté na výber regresorov v lineárnych zmiešaných modeloch pomocou ℓ_1 penalizácie je možné nájsť napríklad v [Bondell et al., 2010] alebo [Fan and Li, 2012]. Oba články rozširujú výber regresorov aj o výber náhodných efektov, čím sa ale obe metódy dosť komplikujú, a teda sa stávajú nevhodné na použitie v prostredí vysoko dimenzionálnych dát. Autori v článku [Jiang et al., 2008] navrhujú metódu na princípe stepwise regresie, ktorá taktiež nie je dostatočne jednoduchá a nie je použiteľná v prostredí vysoko dimenzionálnych dát.

3.1 Kritériá na porovnávanie submodelov v LMM

Po výbere regresorov väčšinou dostaneme na výber z viacerých možností, napríklad pre rôzne λ dostávame zväčša rôzne množiny. V nedávnej dobe sa objavilo mnoho modifikácií informačných kritérií pre LMM, asi najviac pozornosti sa dostalo cAIC - conditional AIC [Vaida and Blanchard, 2005]. Podrobný prehľad a porovnanie prináša práca [Müller et al., 2013]. Napriek všetkému sa ale vo vysoko dimenzionálnych prípadoch najčastejšie stretávame s použitím BIC upraveného pre LMM

$$\mathrm{BIC} = (p + q) \cdot \ln(n) - 2\ln(L).$$

4 Konvexná metóda na výber regresorov vo vysoko dimenzionálnych lineárnych zmiešaných modeloch

V predchádzajúcich kapitole bol zmienený problém existujúcich metód, ktoré optimalizujú vo všeobecnosti nekonvexné účelové funkcie. Pri navrhovaní novej metódy sme sa zamerali hlavne jej použitie na vysoko dimenzionálne dáta s dimenziou väčšou ako 10^4. Za tým účelom sme navrhli konvexnú účelovú funkciu. Nekonvexita účelovej funkcie bola vnášaná do spomínaných metód odhadovaním parametrov od ktorých závisia matice $\mathbf{D}$ a $\mathbf{R}$. Odhad týchto parametrov sme nahradili priamo odhadom vektoru náhodných efektov $\mathbf{u}$. Priame odhadovanie vektora náhodných efektov so sebou ale prináša nevýhodu v podobe toho, že odhadujeme dva vektory ($\boldsymbol{\beta}$ a $\mathbf{u}$) pri ktorých predpokladáme rôznu štruktúru. Kým od vektora $\boldsymbol{\beta}$ očakávame riedku štruktúru, tak vektor $\mathbf{u}$ má normálne rozdelenie. S prihliadnutím na to sme zvolili rôzne formy penalizácie ℓ_1 pre $\boldsymbol{\beta}$ a ℓ_2 pre $\mathbf{u}$. Optimalizovaná účelová funkcia bude mať tvar:

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}) = \arg\min_{\boldsymbol{\beta},\mathbf{u}} \left[\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \Lambda \sum_{i=1}^{q^*} w_i \|_i \mathbf{u}\|_2^2 \right], \tag{4-1}$$

kde q^* je známy počet skupín do ktorých sú rozdelené efekty vo vektore $\mathbf{u}$, ${}_i\mathbf{u}$ je podvektor vektora $\mathbf{u}$ ktoré patrí do i-tej skupiny. λ a Λ sú parametre zvolené pred minimalizáciou a w_i sú zvolené váhy pomocou ktorých budeme škálovať penalizačný parameter Λ pre prislúchajúce časti vektora $\mathbf{u}$.

Častou úlohou analýzy dát pomocou LMM je najmä odhad kovariančnej štruktúry modelu a teda matice $\mathbf{D}$. Ale v našom prípade vysoko dimenzionálnych dát, keď cielime na selekciu regresorov z matice $\mathbf{X}$ nám postačuje dobrý odhad náhodných efektov $\mathbf{u}$ a eliminácia náhodnej časti $\mathbf{Zu}$. K odhadu kovariančnej štruktúri môžeme pristúpiť po selekcii regresorov bežnými postupmi odhadovania LMM.

4.1 Voľba váh w_i

Pri vhodnej voľbe váh w_i v účelovej funkcii vieme 4-1 vieme dosiahnuť dobré výsledky pri malej výpočtovej náročnosti, preto je dôležitá voľba vhodných váh w_i. My navrhujeme:

$$w_i = \frac{1 - {}_i\theta}{{}_iq}, \tag{4-2}$$

kde ${}_iq$ je počet prediktorov v matici $\mathbf{Z}$ prislúchajúcich do i-tej skupiny (počet náhodných efektov v podvektore ${}_i\mathbf{u}$), $q = \sum_{i=1}^{q^*} {}_iq$. ${}_i\theta$ je priemer absolútnych hodnôt korelácie medzi prediktormi z matice $\mathbf{Z}$ prislúchajúce k i-tej skupine (${}_{\mathbf{u}_i}\mathbf{Z}$) a pozorovaniami $\mathbf{Y}$:

$${}_i\theta = \frac{\sum_{i=1}^{iq} |\rho({}_{\mathbf{u}_i}\mathbf{Z}_{(:,i)}, \mathbf{Y})|}{{}_iq}$$

Takáto voľba váh w_i normalizuje subvektor ${}_i\mathbf{u}$ jeho dimenziou a zároveň penalizuje viac časti s menšou priemernou koreláciou, ktorá je pri daných predpokladoch priamo úmerná veľkosti neznámeho efektu.

4.2 Triviálna metóda

Keďže metódy vyvinuté na analýzu vysoko dimenzionálnych v oblasti LMM spomínané v časti 3 nie sú momentálne schopné analyzovať dáta s viac ako niekoľko stoviek regresormi, navrhli sme, pre potreby porovnania, triviálny postup transformácie LMM na model lineárnej regresie. Dáta transformujeme tak aby sme eliminovali náhodnú časť LMM reprezentovanú $\mathbf{Zu}$.

$$\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{Z}\mathbf{Z}^+)\mathbf{X},$$
$$\tilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{Z}\mathbf{Z}^+)\mathbf{Y},$$

kde $\mathbf{Z}^+$ je pseudoinverzná matica k matici $\mathbf{Z}$.

Po transformácii dostaneme lineárny regresný model:

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}},$$

s vysoko dimenzionálnou maticou $\tilde{\mathbf{X}}$. Na takúto úlohu môžeme použiť metódu LASSO z časti 2. V tomto prípade môžu nastať rôzne problémy, ktoré môžu skresliť selekciu premenných pomocou LASSA. Napríklad v prípade zhody (vysokej korelácie) medzi stĺpcami z matíc $\mathbf{X}$ a $\mathbf{Z}$. Zároveň, týmto postupom nemôžeme analyzovať dáta, v prípade ak je počet prediktorov väčší ako počet pozorovaní (teda musí platiť $q \ll n \ll p$).

5 Teoretické vlastnosti navrhnutých metód

V tejto časti rozoberieme schopnosť navrhnutej metódy identifikovať presnú množinu premenných, ktoré ovplyvňujú pozorovania. V prípade vysoko dimenzionálnych dát sa v literatúre stretávame najmä s analýzou asymptotických vlastností. V prácach [Zhao and Yu, 2006, Knight and Fu, 2000] sa stretávame najmä s konzistenciou odhadu a konzistencie z pohľadu výberu premenných:

$$P(\hat{S}^n = S^0) \to 1, \text{ pre } n \to \infty, \tag{5-1}$$

kde $\hat{S}^n$ je množina odhadnutá pomocou danej metódy. V našom prípade sa pozrieme na **znamienkovú konzistenciu** a to:

$$\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}^n(\lambda^n,\Lambda) =_s \boldsymbol{\beta}^0) = 1, \tag{5-2}$$

kde $\hat{\boldsymbol{\beta}}^n(\lambda^n,\Lambda) =_s \boldsymbol{\beta}^0$ znamená $\text{sign}(\hat{\boldsymbol{\beta}}^n(\lambda^n,\Lambda)) = \text{sign}(\boldsymbol{\beta}^0)$. Táto vlastnosť je prísnejšia ako konzistencia lebo okrem správneho výberu premenných zaručuje aj správnosť znamienka.

Bez ujmy na všeobecnosti predpokladajme, že prvých k premenných je relevantných. Rozdeľme vektor pevných efektov $\boldsymbol{\beta}^0$ na $\boldsymbol{\beta}^0(1) = (\beta_1^0, \beta_2^0, \ldots, \beta_k^0)$, kde $\beta_j^0 \neq 0$ pre $j = 1,2,\ldots,k$ a $\boldsymbol{\beta}^0(2) = (\beta_{k+1}^0, \beta_{k+2}^0, \ldots, \beta_p^0)$, kde $\beta_j^0 = 0$ pre $j = k+1, k+2, \ldots, p$. Prislúchajúc k tomu rozdelme aj maticu $\mathbf{X}^n$ na $\mathbf{X}^n(1)$ a $\mathbf{X}^n(2)$. A definujme:

$$\boldsymbol{\Sigma}^n = \frac{1}{n}[\mathbf{X}^n, \mathbf{Z}^n]^T[\mathbf{X}^n, \mathbf{Z}^n] = \begin{pmatrix} \boldsymbol{\Sigma}_{1,1}^n & \boldsymbol{\Sigma}_{1,2}^n & \boldsymbol{\Sigma}_{1,3}^n \\ \boldsymbol{\Sigma}_{2,1}^n & \boldsymbol{\Sigma}_{2,2}^n & \boldsymbol{\Sigma}_{2,3}^n \\ \boldsymbol{\Sigma}_{3,1}^n & \boldsymbol{\Sigma}_{3,2}^n & \boldsymbol{\Sigma}_{3,3}^n \end{pmatrix}$$

kde $\boldsymbol{\Sigma}_{1,1}^n = \frac{1}{n}(\mathbf{X}^n(1))^T\mathbf{X}^n(1)$, $\boldsymbol{\Sigma}_{1,2}^n = \frac{1}{n}(\mathbf{X}^n(1))^T\mathbf{X}^n(2)$, $\boldsymbol{\Sigma}_{1,3}^n = \frac{1}{n}(\mathbf{X}^n(1))^T\mathbf{Z}^n$, $\boldsymbol{\Sigma}_{2,1}^n = \frac{1}{n}(\mathbf{X}^n(2))^T\mathbf{X}^n(1)$, $\boldsymbol{\Sigma}_{2,2}^n = \frac{1}{n}(\mathbf{X}^n(2))^T\mathbf{X}^n(2)$, $\boldsymbol{\Sigma}_{2,3}^n = \frac{1}{n}(\mathbf{X}^n(2))^T\mathbf{Z}^n$, $\boldsymbol{\Sigma}_{3,1}^n = \frac{1}{n}(\mathbf{Z}^n)^T\mathbf{X}^n(1)$, $\boldsymbol{\Sigma}_{3,2}^n = \frac{1}{n}(\mathbf{Z}^n)^T\mathbf{X}^n(2)$, $\boldsymbol{\Sigma}_{3,3}^n = \frac{1}{n}(\mathbf{Z}^n)^T\mathbf{Z}^n$. Na základe toho definujme podmatice

$$\boldsymbol{\Psi}^n = \begin{pmatrix} \boldsymbol{\Sigma}_{1,1}^n & \boldsymbol{\Sigma}_{1,3}^n \\ \boldsymbol{\Sigma}_{3,1}^n & \boldsymbol{\Sigma}_{3,3}^n + \frac{\Lambda}{n}\mathbf{I} \end{pmatrix} \quad \text{a} \quad \boldsymbol{\Delta}^n = \begin{pmatrix} \boldsymbol{\Sigma}_{2,1}^n & \boldsymbol{\Sigma}_{2,3}^n \end{pmatrix}.$$

Ďalej definujme vektor

$$\boldsymbol{\theta} = \begin{pmatrix} \text{sign}(\boldsymbol{\beta}^0(1)) \\ \mathbf{0}_{q\times 1} \end{pmatrix}.$$

Definícia 1. Hovoríme, že $\mathbf{\Sigma}^n$ spĺňa **podmienku nereprezentovateľnosti** (pre naše potreby upravená Irrepresentable condition, ktorá sa vyskytuje v literatúre študujúcej LASSO) ak existuje vektor kladných konštánt $\boldsymbol{\eta}$ pre ktorý platí:

$$\left|\mathbf{\Delta}^n(\mathbf{\Psi}^n)^{-1}\boldsymbol{\theta}\right| < 1 - \boldsymbol{\eta}$$

kde $\mathbf{1}$ je vektor $p - k + q$ jednotiek.

Ako uvádza [Bühlmann and van de Geer, 2011, Zhao and Yu, 2006] podmienka nereprezentovateľnosti je vo všeobecnosti ekvivalentná s podmienkou, ktorá je v literatúre známa ako podmienka susedskej stability (neighborhood stability condition) a obe zlyhávajú ak dizajn matíc $\mathbf{X}$ a $\mathbf{Z}$ vykazuje až príliš silnú lineárnu závislosť v menšej podmatici matíc $\mathbf{X}$ a $\mathbf{Z}$.

Najskôr stanovíme dolnú hranicu pravdepodobnosti, s ktorou naša metóda vyberie správnu podmnožinu premenných, čo je tesne prepojené s tým „ako veľmi platí“ podmienka nereprezentovateľnosti.

Lemma 1. *Ak platí podmienka nereprezentovateľnosti, teda existuje vektor kladných konštánt* $\boldsymbol{\eta}$ *potom:*

$$P(\hat{\boldsymbol{\beta}}^n(\lambda^n,\Lambda) =_s \boldsymbol{\beta}^0) \geq P(A^n \cap B^n),$$

kde

$$A^n = \left\{ \left|(\mathbf{\Psi}^n)^{-1}\mathbf{\Phi}^n\boldsymbol{\xi}^n\right| < n\,|\boldsymbol{w}| - \frac{\lambda^n}{2}\left|(\mathbf{\Psi}^n)^{-1}\boldsymbol{\theta}\right| \right\}$$

a

$$B^n = \left\{ \left|(\mathbf{\Delta}^n(\mathbf{\Psi}^n)^{-1}\mathbf{\Phi}^n - (\mathbf{X}^n(2))^T)\boldsymbol{\xi}^n\right| \leq \frac{\lambda^n}{2}\boldsymbol{\eta} \right\}.$$

kde

$$\mathbf{\Phi}^n = \begin{pmatrix} (\mathbf{X}^n(1))^T \\ (\mathbf{Z}^n)^T \end{pmatrix}.$$

Dôkaz lemmy 1 ako aj dôkaz nasledujúcej vety 1 je uvedený v dizertačnej práci.

Veta 1. *Naša metóda je znamienkovo konzistentná pre postupnosť* λ_n *spĺňajúcu* $\lambda_n/n \to 0$ *a* $\lambda_n/n^{\frac{1+c}{2}} \to \infty$, *kde* $0 \leq c < 1$ *za predpokladu konečných variančných matíc* χ_1 *a* χ_2 *a platiacej podmienka nereprezentovateľnosti. Zároveň platí:*

$$P(\hat{\boldsymbol{\beta}}(\lambda_n) =_s \boldsymbol{\beta}^0) = 1 - o(e^{-n^c})$$

6 Príklady použitia navrhnutých metód

V tejto časti odprezentujeme použitie navrhnutej metódy na konkrétnych príkladoch. Prejdeme základné vygenerované príklady, kde odhady (parametrov) bude môcť porovnať so skutočnosťou. Ďalej uvedieme aj postup v situácii, keď je dimenzia matice $\mathbf{Z}$, q, väčšia ako počet pozorovaní n a v závere sa pozrieme na aplikáciu navrhnutej metódy na reálne dáta.

6.1 Základný príklad

Začneme vygenerovaným príkladom v ktorom budeme poznať skutočnosť, aby sme si s ňou mohli výsledky porovnať.

Uvažujme jednoduchý LMM so $n = 200$ pozorovaniami. Počet stĺpcov matice $\mathbf{X}$ (počet regresorov) je $p = 1000$. Z týchto je len prvých $k = 5$ významných pre daný model s pevným efektom 1. Vektor pevných efektov teda je $\boldsymbol{\beta} = [1, 1, 1, 1, 1, 0, \ldots, 0]^T$. Počet prediktorov v matici $\mathbf{Z}$ je $q = 40$. Matica $\mathbf{Z}$ je blokovo diagonálna. Náhodné efekty sú rozdelené do dvoch skupín po 20 efektov. Prvá skupina efektov je z normálneho rozdelenia s nulovou strednou hodnotou a varianciou $\sigma_1^2 = 1{,}2$, druhá skupina má varianciu $\sigma_2^2 = 0{,}8$. LMM má dva variančné komponenty $\mathbf{D} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$. Chyby $\boldsymbol{\varepsilon}$ sú z normálneho rozdelenia s nulovou strednou hodnotou a varianciou $\sigma_\varepsilon^2 = 0.2$. Matice $\mathbf{X}$ a $\mathbf{Z}$ sú generované z rovnomerného rozdelenia rozdelenia a normované aby mal každý stĺpec dĺžku jedna. Pozorovania $\mathbf{Y}$ sú vypočítané na základe LMM:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\,\mathbf{u} + \boldsymbol{\varepsilon}.$$

Na danú úlohu aplikujeme navrhnutú metódu pre rôzne parametre λ a Λ. Váhy w_i sme volili podľa 4.1. Zároveň sme pre každú kombináciu parametrov λ a Λ a teda pre každú podmnožiny vybratých regresorov odhadli LMM.

Na základe BIC sa môžeme rozhodnúť, ktorý z ponúkaných modelov si zvolíme. V našom prípade sme zvolili model v ktorom je odhad efektov prvých piatich relevantných regresorov $(0.85, 1.02, 0.82, 0.96, 0.83)$ a je nesprávne identifikovaných päť ďalších regresorov, ktorých efekty sú odhadnuté na $(0.12, 0.15, 0.09, 0.1, 0{,}08)$. V tomto prípade je odhad variančných efektov (v zátvorkách sú skutočné hodnoty) $1.05(1.2)$ a $0.33(0.8)$. Odhad variancie chýb je $0.02(0.2)$.

Keď použijeme presný model len s prvými piatimi regresormi tak odhad pevných efektov je $(0.93, 1.1, 0.92, 1.01, 0.93)$. Odhady variančných efektov sú 1.05, 0.33 a odhad variancie chýb je 0.04.

6.2 Analýza dát z Wellcome Trust Case Control Consortium (WTCCC)

LMM sa často využívajú na mapovanie biologických znakov na základe genetickej informácie [Zhou et al., 2013, Tan et al., 2018]. Tento postup sa v literatúre označuje ako genome-wide association (GWA) štúdie. Sú založené na tom, že akékoľvek dva ľudské genómy sa líšia v miliónoch rôznych spôsobov. Existujú malé rozdiely v jednotlivých nukleotidoch genómov (SNP), ako aj veľa väčších variácií, ako sú delécie, inzercie a variácie počtu kópií. Každá z nich môže spôsobiť zmeny v znakoch jednotlivca alebo fenotypu, čo môže byť čokoľvek, od rizika choroby až po fyzické vlastnosti, ako je výška.

Úloha spočíva v regresii kvantitatívnych alebo binárnych znakov na vysoko dimenzionálny vektor (v prípade WTCCC dát 450000) jednonukleotidových polymorfizmov (SNP), alebo lokácii diskrétnych genetických variácií pre každého jednotlivca v štúdii. Náhodné efekty môžu byť

Tabuľka 6 – 1 Porovnanie metód BSLMM, arLMM-AVC a LMMconvexLASSO pri predikcii dát z WTCCC.

choroba	AUC		
	LMMconvexLASSO	arLMM-AVC	BSLMM
BD	0.6687	0.6461	0.6514
CAD	0.6172	0.5937	0.6015
HT	0.6164	0.6010	0.5893
RA	0.6053	0.6206	0.6178
T1D	0.6892	0.6840	0.6932
T2D	0.6152	0.5993	0.6015

spôsobené štruktúrou obyvateľstva alebo príbuznosťou medzi jednotlivcami. V našom prípade budeme uvažovať (rovnako ako v zmienených článkoch populačnú štruktúru v tvare $\mathbf{X}\mathbf{X}^{\mathrm{T}}$)

Porovnávame náš odhad metódou LMMconvexLASSO s BSLMM (Bayesian sparse linear mixed model) [Zhou et al., 2013] a s arLMM-AVC (Approximate Ridge LMM - approximate variance componens) [Tan et al., 2018]. Na porovnanie použijeme dáta zo štúdie Wellcome Trust Case Control Consortium (WTCCC) [Consortium et al., 2007] s 14 000 chorých pacientov pre 6 ochorení - bipolárna porucha (BD), ochorenie koronárnych artérií (CAD), hypertenzia (HT), reumatoidná artritída (RA), diabetes typu 1 (T1D) a diabetes typu 2 (T2D) a s 3 000 zdieľanými zdravými. Predikujeme binárnu premennú - zdravotný stav (chorý, zdravý).

Náhodne rozdelíme dáta na trénovaciu časť (80 % dát) a testovaciu časť (zostávajúcich 20 % dát). Tento proces zopakujeme 20 krát pre každú chorobu. Pre každú metódu odhadneme parametre modelu na trénovacej časti a následne predikujeme na testovacej časti. Výsledky vyhodnotené pomocou metódy založenej na porovnaní plochy pod ROC krivkou (AUC - area under the curve) [Wray et al., 2010] vidíme v 6 -- 1. ROC krivka je grafickým znázornením závislosti senzitivity od 1 - špecificity v závislosti od prahu zvoleného klasifikátora. Plocha pod ROC krivkou (AUC) je jedním z kritérií pre výber vhodného klasifikátoru (modelu) a určuje kvalitu separácie tried v modeli reps. schopnosť separácie tried daným klasifikátorom.

V tabuľke 6 -- 1 vidíme, že úspešnosť jednotlivých metód pri predikcii chorôb je takmer totožná. Zároveň ale vidno, že v danej oblasti určite existuje priestor na ďalšie vylepšenie metód, lebo maximálna úspešnosť je necelých 0.7. Potvrdzuje sa, že predikcia choroby z genetických markerov je ťažký problém lebo veľké rozdiely v odozve môžu vzikať environmentálnymi faktormi, ktoré v uvažovanej štúdii nepozorujeme.

7 Simulačná štúdia

V tejto kapitole odprezentujeme výsledky viacerých simulačných štúdií, v ktorých sa postupne zameriame na rôzne časti LMM od rastúceho počtu regresorov až po komplikovanosť kovariančnej matice náhodných efektov $\mathbf{u}$. Zároveň uvedieme porovnanie navrhnutej metódy s existujúcimi metódami popísanými v 3. Takmer vždy použijeme na porovnanie nami navrhnú metódu a metódu navrhnutú na porovnávanie popísanú v časti 4.2.

Výsledkom každej z metód popísaných v tejto práci je odhad pevných efektov $\boldsymbol{\beta}$. My sa ale

budeme zameriavať na výber regresorov a preto ako množinu vybratých regresorov pomocou danej metódy budeme uvažovať tie regresori, ktorých odhad pevných efektov pomocou danej metódy je nenulový. Teda vyberieme tú množinu regresorov, pre ktorú platí:

$$\{i : \hat{\beta}_i^* \neq 0 \quad \text{pre} \quad i = 1, 2, \ldots, p\}$$

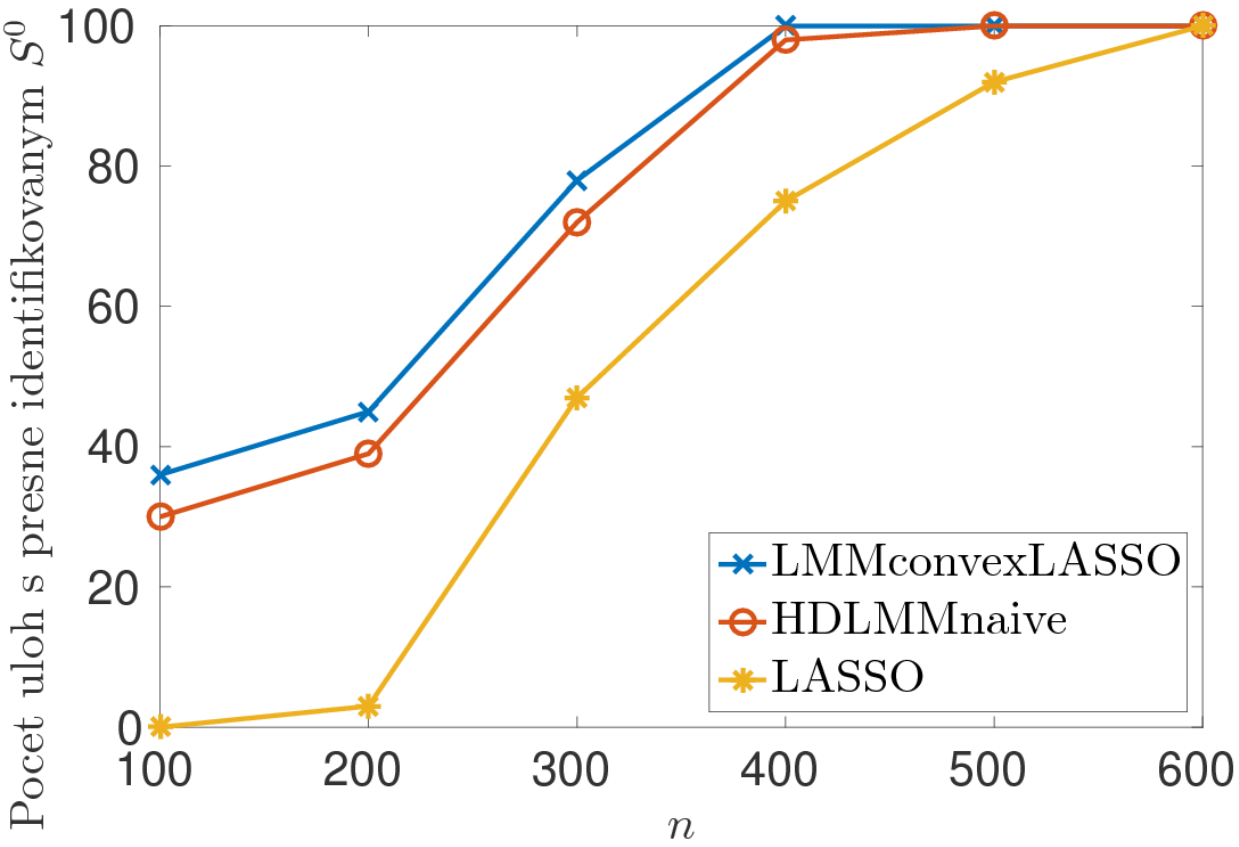
kde $\hat{\beta}_i^*$ je odhad pomocou danej metódy.

V každej štúdii vždy pre každý skúmanú časť vygenerujeme sto úloh podľa popísaného postupu a použijeme všetky porovnávané metódy. Pre danú metódu budeme sledovať v koľkých zo sto prípadov nám daná metóda poskytne aspoň pre jeden parameter alebo skupinu parametrov, ktoré sa v metóde menia (napríklad v LASSO metóde λ) práve presnú množinu významných regresorov, teda práve množinu S^0. V každej štúdii porovnáme aspoň tri metódy a to nami navrhnutú metódu z časti 4 s metódou navrhnutou na porovnávanie z časti 4.2 a s klasickou LASSO metódou z časti 3, ktorá je navrhnutá len pre klasický lineárny regresný model a nie pre LMM. Zároveň, často budeme porovnávať len tieto tri metódy, lebo ako jediné sú schopné efektívne počítať úlohy s dimenziou väčšou ako 10^3.

Rastúci počet pozorovaní

Najskôr sa pozrieme na zmenu úspešnosti nájdenia práve množiny S^0 s rastúcim počtom pozorovaní n. Uvažujme LMM postupne s $n = 100, 200, 300, 400, 500,$ a 1000 pozorovaniami. Nech $p = 5000$, $q = 40$ a $\mathbf{D}$ je diagonálna s dvoma varinčnými komponentami $\sigma_1^2 = 1.2$ a $\sigma_2^2 = 0.8$, $\sigma_\varepsilon^2 = 0.2$. Nech prvých $s^0 = 15$ regresorov je významných s efektom jedna. Obe matice $\mathbf{X}$ aj $\mathbf{Z}$ sú generované z rovnomerného rozdelenia a znormované tak aby mal každý stĺpec jednotkovú dĺžku.

Nasledujúci obrázok 7 -- 2 potvrdzuje očakávania, a to že s rastúcim počtom pozorovaní narastá úspešnosť každej z metód.

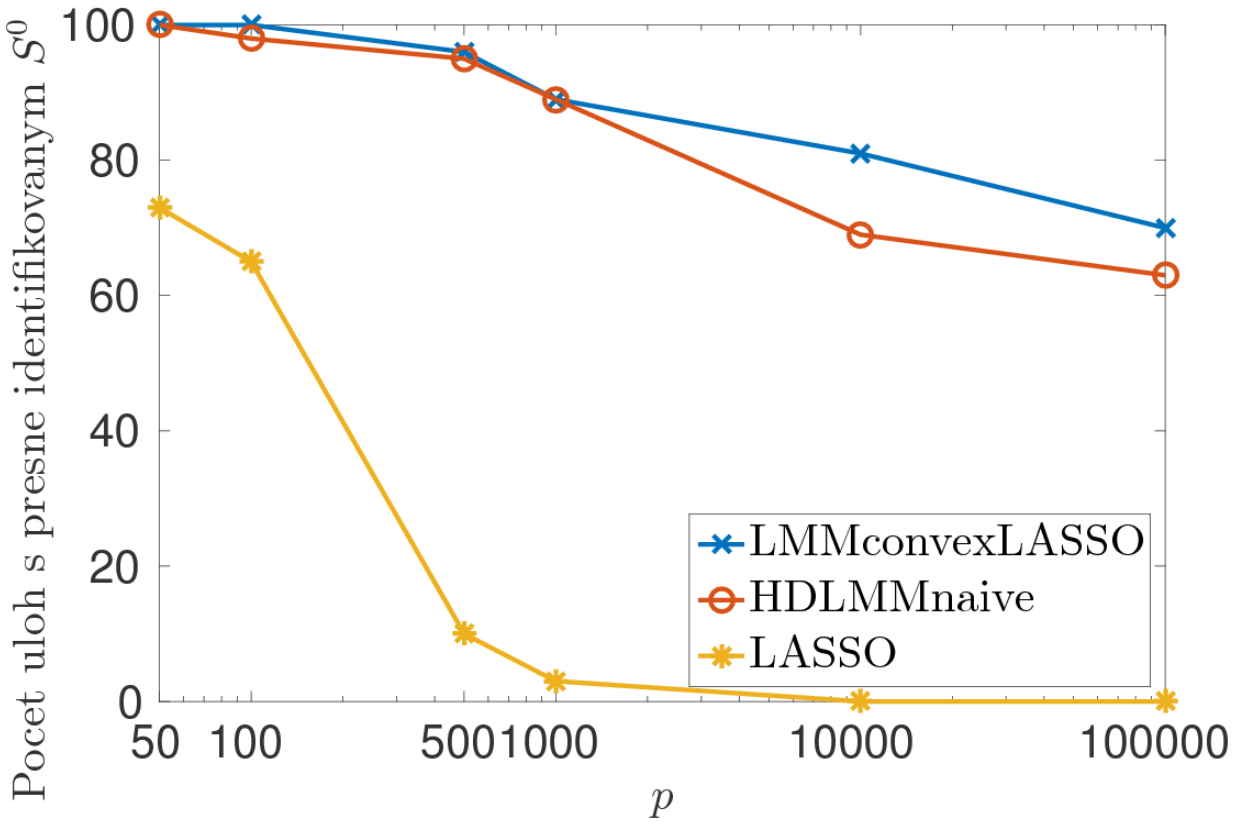


Obr. 7 – 2 Porovnanie úspešnosti metód v závislosti od počtu pozorovaní.

Rastúci počet regersorov v matici X

Ďalej sa pozrieme, ako sa zmení úspešnosť nájdenia páve množiny S^0 s narastajúcou dimenziou p matice $\mathbf{X}$. Uvažujme LMM, v ktorom bude mať matica $\mathbf{X}$ postupne $p = 50, 100, 500, 1000, 10000, 100000$ regresorov. Vždy bude významných práve $s^0 = 5$ regresorov, ktorých efekt bude jedna. Regresory budú vyberané náhodne z rovnomerného rozdelenia a normované na jednotkovú dĺžku. Ďalšie parametre sú $n = 200$, $q = 40$ a $\mathbf{D}$ je diagonálna s dvoma varinčnými komponentami $\sigma_1^2 = 1.2$ a $\sigma_2^2 = 0.8$, $\sigma_\varepsilon^2 = 0.2$. Obe matice $\mathbf{X}$ aj $\mathbf{Z}$ sú generované z rovnomerného rozdelenia a znormované tak aby mal každý stĺpec jednotkovú dĺžku.

Ako bolo spomenuté v časti 2 s rastúcim počtom regresorov narastá pravdepodobnosť, že sa náhodne vyskytnú regresory, ktorých vysoká korelácia s vektorom pozorovaní $\mathbf{Y}$ bude čisto náhodná. Obrázok 7 -- 3 ukazuje pokles úspešnosti s rastúcim počtom regresorov. V porovnaní s nasledujúcim obrázkom 7 -- 4 je pokles úspešnosti metód spojený s nárastom počtu regresorov výrazne menší ako pokles úspešnosti spojený s rastom počtu prvkov množiny S^0, obrázok 7 -- 4.

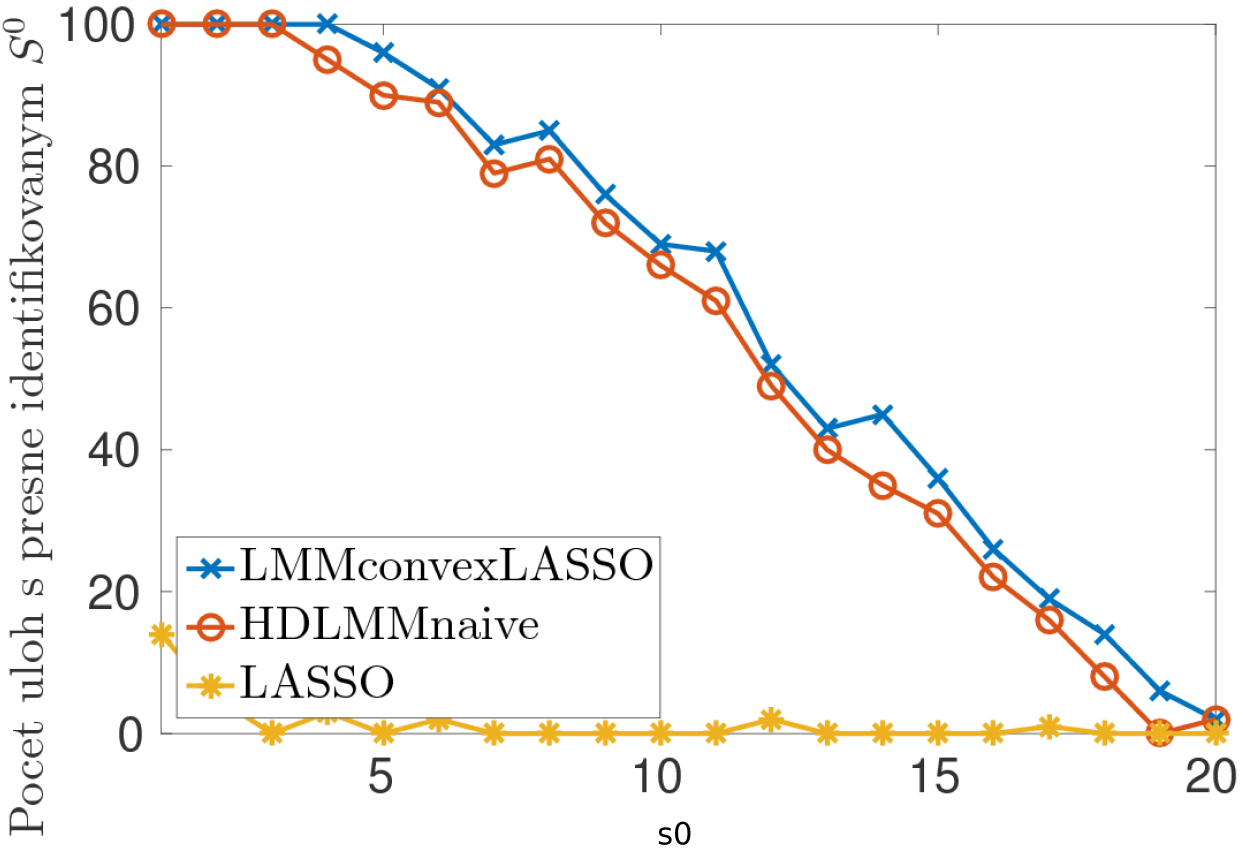


Obr. 7 – 3 Porovnanie úspešnosti metód v závislosti od počtu regresorov.

Rastúci počet relevantných regersorov s^0

Teraz sa pozrieme ako sa zmení úspešnosť metód pri náraste počtu relevantných regresorov. Uvažujme postupne LMM s $s^0 = 1,2,\ldots,20$ relevantnými regresormi. Všetky regresori budú mať efekt jedna. Ďalšie parametre LMM sú $p = 5000$, $n = 200$, $q = 40$ a $\mathbf{D}$ je diagonálna s dvoma varinčnými komponentami $\sigma_1^2 = 1.2$ a $\sigma_2^2 = 0.8$, $\sigma_\varepsilon^2 = 0.2$. Obe matice $\mathbf{X}$ aj $\mathbf{Z}$ sú generované z rovnomerného rozdelenia a znormované tak aby mal každý stĺpec jednotkovú dĺžku.

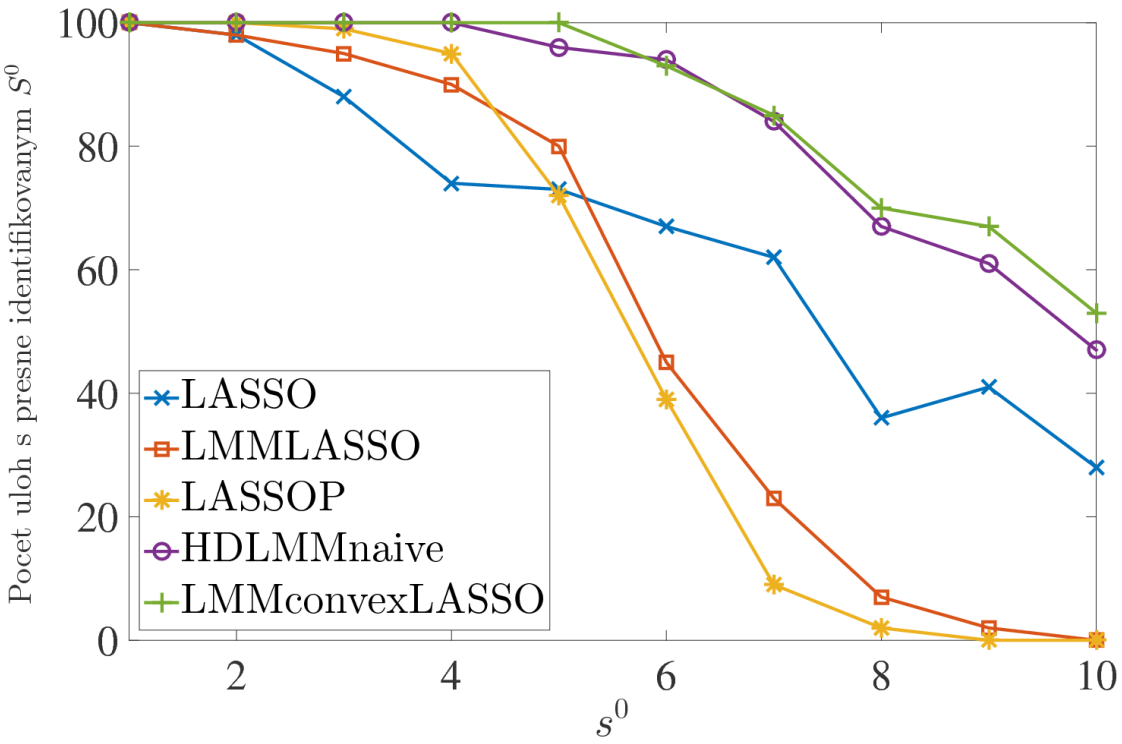
S narastajúcim počtom relevantných regresorov je zrejmý relatívny pokles vplyv jednotlivých regresorov na pozorovania $\mathbf{Y}$, čomu zodpovedá aj postupný pokles úspešností jednotlivých metód ako to vidieť na nasledujúcom obrázku 7 -- 4.



Obr. 7 – 4 Porovnanie úspešnosti metód v závislosti od veľkosti množiny S^0.

Porovnanie s existujúcimi metódami

Na porovnanie navrhovaných metód s už existujúcimi metódami, výrazne zmenšiť dimenziu riešených úloh. Najskôr porovnáme metódy s metódami LASSOP a LMMLASSO (časť 3). Metódy porovnávame na príkladoch, ktoré sú odvodené z ilustračných príkladov k daným metódam. Majme $\mathfrak{n} = 120$ pozorovaní rozdelených do dvadsiatich skupín po šesť. Pre každé pozorovanie máme $\mathfrak{p} = 150$ regresorov. Postupne nech je len $s^0 = 1, 2, \ldots, 10$ regresorov významných s efektom jedna. Matica $\mathbf{Z}$ zachytáva rozdelenie dát do skupín. Pre každú skupinu pozorujeme dva nezávisle prediktory. Matica $\mathbf{Z}$ je blokovo diagonálna a vektor $\mathfrak{u}$ pozostáva z dvoch častí. Obe časti vektora $\mathfrak{u}$ sú náhodne vybraté z rozdelenia $\mathcal{N}(0, \mathbf{D} = 2 \cdot \mathbf{I})$, chyby sú z $\mathcal{N}(0,\mathbf{I})$.

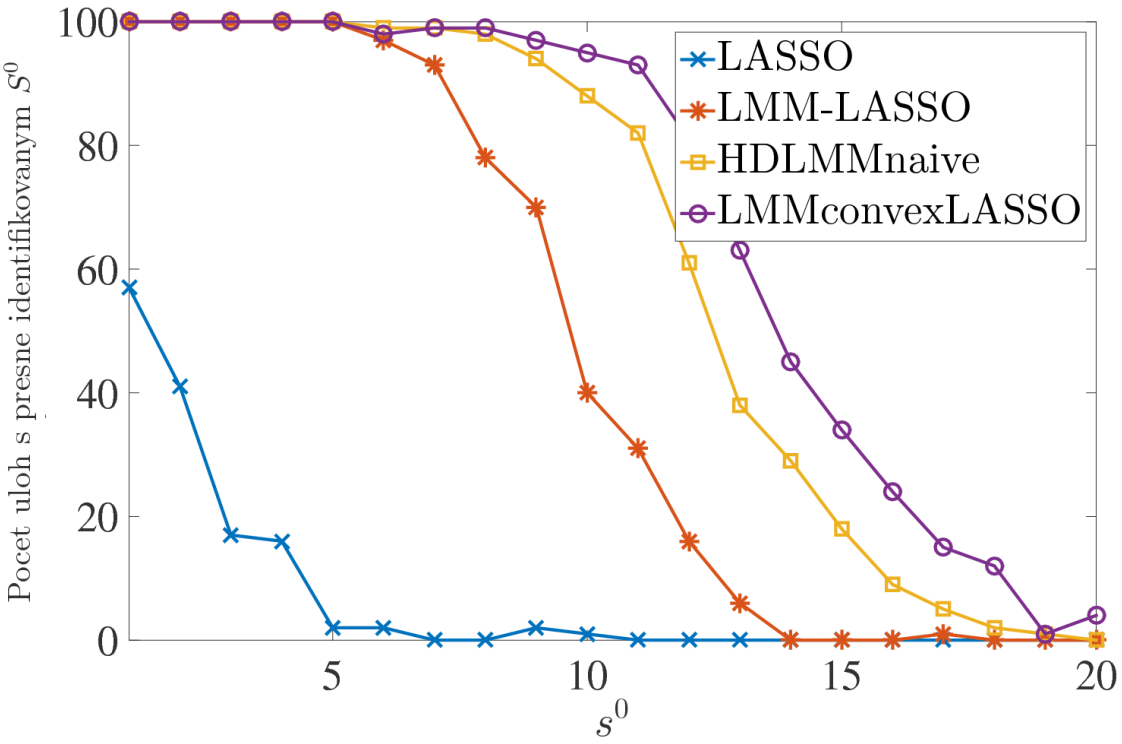


Obr. 7 – 5 Porovnanie úspešnosti metód v závislosti od veľkosti množiny S^0.

Vidno , že obe metódy (LASSOP aj LMMLASSO) majú veľmi podobné výsledky, čo zodpovedá aj tvrdeniu autorov. Zároveň vidieť, že obe nami navrhnuté metódy sú úspešnejšie ako porovnávané metódy. Zarážajúci je aj fakt, že s rastúcim počtom relevantných regresorov rapídne klesá úspešnosť daných metód až pod úroveň úspešnosti LASSO metódy, ktorá nie je navrhnutá pre LMM.

V druhom prípade porovnáme navrhnuté metódy s metódou LMM-LASSO (časť 3), ktorá je navrhnutá na problémy s jedným variačným komponentom.

V tomto prípade budeme mať $\mathbf{n} = 200$ pozorovaní rozdelených do dvadsiatich skupín po desiatich pozorovaniach. pre každé pozorovanie máme $\mathbf{p} = 5000$ regresorov, ale z nich je len $s^0 = 1, 2, \ldots, 20$, všetky s efektom jedna. Matica $\mathbf{Z}$ zachytáva skupinovú štruktúru dát. $\mathbf{Z}_{i,j}$ je jedna ak pozorovania $\mathbf{i}$ a $\mathbf{j}$ patria do rovnakej skupiny a nula ak nie. Vektor $\mathbf{u}$ je náhodne vybratý z rozdelenia $\mathcal{N}(0, \mathbf{D} = \mathbf{I})$, chyby sú z $\mathcal{N}(0{,}0.2\mathbf{I})$.



Obr. 7 – 6 Porovnanie úspešnosti metód v závislosti od veľkosti množiny S^0.

LMM-LASSO metóda si vedie porovnateľne s nami navrhnutými metódami. Problémy môžu nastať v prípade ak by boli regresory príliš korelované s prediktormi, lebo metóda najskôr odhaduje kovariačnú štruktúru modelu, ktorú by mohlo v takomto prípade výrazne nadhodnotiť.

Ďalšie simulačné štúdie je možné nájsť v dizertačnej práci.

Zdrojové kódy

Všetky navrhnuté algoritmy sme implementovali v prostredí `MATLAB`. Na minimalizáciu sme použili prostredie CVX: Matlab Software for Disciplined Convex Programming [Grant and Boyd,] s optimalizačným algoritmom MOSEK [Mosek ApS,]. Zdrojové kódy s ukážkami sú k dispozícii na `http://www.mathworks.com/matlabcentral/fileexchange/56952-lmmconvexlasso`.

Záver

V práci sme v úvodných častiach 2 sme sa oboznámili s vysoko dimenzionálnymi dátami a ich problematikou v prípade lineárnej regresie. Uviedli sme prehľad základných metód používaných v tejto oblasti. Ďalej sme zadefinovali lineárny zmiešaný model v klasickej podobe a uviedli zaužívané postupy odhadovania LMM. Nasledovne sme uviedli prehľad prác ktoré boli v posledných rokoch publikované v oblasti vysoko dimenzionálnych lineárnych zmiešaných modelov. Poukázali sme najmä na fakt, že aj keď sú tieto metódy navrhnuté na analýzu vysoko dimenzionálnych dát nie sú vďaka nekonvexite účelovej funkcie vhodné na analýzu dát, ktorých dimenzia je často vyššia ako 10^3.

V nasledujúcej časti 3 sme sa zamerali na výber regresorov vo vysoko dimenzionálnych lineárnych zmiešaných modeloch. Za týmto účelom sme navrhli konvexnú metódu, ktorú sme postupne vylepšovali až do metódy 4-1. Táto metóda využíva rôzne typy penalizácií a váh pre rôzne časti vektora odhadovaných parametrov. Následne sme navrhli voľbu vhodných váh. Zároveň sme navrhli jednu triviálnu metódu. Výhodou navrhnutých metód je že dokážu riešiť problémy s počtom premenných presahujúci 10^5. Možnou nevýhodou je že pre úspešný odhad požadujeme aby bol počet pozorovaní väčší ako počet parametrov ktoré sa snažíme vybrať/odhadnúť. V časti 4 ale navrhujeme postup, ktorý by mohlo byť možné uplatniť v takýchto prípadoch.

V nasledujúcej kapitole 5 sme odvodili teoretické vlastnosti navrhnutej metódy. Konkrétne konzistenciu ktorá zabezpečí, že s rastúcim počtom pozorovaní, za daných predpokladov, určite vyberieme požadovanú množinu regresorov. V nasledujúcej kapitole 6 sme uviedli príklady použitia navrhnutej metódy najskôr na simulovaných a potom na reálnych dátach. V tejto kapitole sme taktiež navrhli postup, aplikovateľný v prípadoch, keď je počet pozorovaní menší ako počet prediktorov.

Posledná kapitola 7 je venovaná simulačnej štúdii. Hlavnou výhodou navrhnutých metód je ich možné použitie aj v problémoch s dimenziou väčšou ako 10^5 ako je možné vidieť na obrázku 7 -- 3.

V porovnaní s existujúcimi metódami sú navrhnuté metódy úspešnejšie, ako je možné vidieť na obrázkoch 7 -- 5 a 7 -- 6, čo pripisujeme najmä lepšej implementácii. Lepšia implementácia je dosiahnutá najmä vďaka použitiu profesionálneho optimalizačného programu určeného na riešenie úloh konvexného programovania.

Výsledky dizertačnej práce

Za najdôležitejšie prínosy dizertačnej práce považujeme:

- Odvodenie konvexnej metódy na výber regresorov vo vysoko dimenzionálnych lineárnych zmiešaných modeloch, pomocou ktorej je možné analyzovať dáta s dimenziou, presahujúcou dimenziu s ktorou sú schopné pracovať doterajšie metódy.
- Návrh váh pre odhad náhodných efektov. Navrhnutá voľba váh výrazne zjednodušuje výpočetnú zložitosť, ale zároveň, ako ukazujú simulácie, má minimálny vplyv na úspešnosť selekcie našej metódy.
- Rozšírenie teoretických poznatkov v oblasti výberu premenných v lineárnych zmiešaných modeloch. Známe teoretické poznatky z tejto oblasti sme rozšírili pre metódu navrhnutú v praci. Konkétne sa jedná o asymptotickú znamienkovú konzistenciu, ktorá zabezpečí, že naša metóda s dostatočným počtom dát vyberie do modela správne premenné.
- Implementácia metódy v programovacom jazyku `MATLAB`. Implementácie nie je závislá od žiadneho ad hoc algoritmu, ale je plne postavená na konvexnom programovaní a teda škáloveteľná s rozvojom konvexného programovania.

Využitie výsledkov v praxi:

- Implementácia metódy je voľne dostupná v v programovaciom jazyku `MATLAB`. Ako bolo uvedené v práci, može byť využitá pri genome-wide association štúdiách, kde dosahuje porovnateľné výsledky s metódami dnes považovanými za „state of the art“.
- Zároveň je možné metódu vyžiť pri riešení podobných problémov v meteorológii alebo finančníctve.

Literatúra

[Atwell et al., 2010] Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J. R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Nemri, A., Roux, F., Salt, D. E., Tang, C., Todesco, M., Traw, M. B., Weigel, D., Marjoram, P., Borevitz, J. O., Bergelson, J. and Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature *465*, 627--631.

[Bai and Ng, 2008] Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. Journal of Econometrics *146*, 304--317.

[Bondell et al., 2010] Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. Biometrics *66*, 1069--1077.

[Bühlmann and van de Geer, 2011] Bühlmann, P. and van de Geer, S. (2011). Statistics for high-dimensional data. Springer.

[Consortium et al., 2007] Consortium, W. T. C. C. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661.

[Eisenhart, 1947] Eisenhart, C. (1947). The assumptions underlying the analysis of variance. Biometrics *3*, 1--21.

[Fan and Lv, 2010] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica *20*, 101.

[Fan and Li, 2012] Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. Annals of statistics *40*, 2043.

[Friedman et al., 2009] Friedman, J., Hastie, T. and Tibshirani, R. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer-Verlag New York.

[Friedman et al., 2010] Friedman, J., Hastie, T. and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736.

[Grant and Boyd,] Grant, M. C. and Boyd, S. P. (.). CVX: Matlab Software for Disciplined Convex Programming. `http://cvxr.com/about/`. Accessed: 2015-09-30.

[Jiang et al., 2008] Jiang, J., Rao, J. S., Gu, Z., Nguyen, T. et al. (2008). Fence methods for mixed model selection. The Annals of Statistics *36*, 1669--1692.

[Knight and Fu, 2000] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. Annals of statistics ., 1356--1378.

[Kushch et al., 2008] Kushch, I., Arendacká, B., Štolc, S., Mochalski, P., Filipiak, W., Schwarz, K., Schwentner, L., Schmid, A., Dzien, A., Lechleitner, M. et al. (2008). Breath isoprene--aspects of normal physiology related to age, gender and cholesterol profile as determined in a proton transfer reaction mass spectrometry study. Clinical chemistry and laboratory medicine *46*, 1011--1018.

[Lippert, 2013] Lippert, C. (2013). Linear mixed models for genome-wide association studies. `https://publikationen.uni-tuebingen.de/xmlui/handle/10900/50003`.

[Miller, 2002] Miller, A. (2002). Subset Selection in Regression, vol. 95, of C&H/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC.

[Mosek ApS,] Mosek ApS (.). High performance software for large-scale LP, QP, SOCP, SDP and MIP including interfaces to C, Java, MATLAB, .NET, R and Python. `https://www.mosek.com/`. Accessed: 2010-09-30.

[Müller et al., 2013] Müller, S., Scealy, J. L. and Welsh, A. H. (2013). Model selection in linear mixed models. Statistical Science *28*, 135--167.

[Pampuri et al., 2011] Pampuri, S., Schirru, A., Fazio, G. and De Nicolao, G. (2011). Multilevel lasso applied to virtual metrology in semiconductor manufacturing. In Automation Science and Engineering (CASE), 2011 IEEE Conference on pp. 244--249, IEEE.

[Rohart et al., 2014] Rohart, F., San Cristobal, M. and Laurent, B. (2014). Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. Computational Statistics & Data Analysis *80*, 209--222.

[Schelldorfer et al., 2011] Schelldorfer, J., Bühlmann, P. and van De Geer, S. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization. Scandinavian Journal of Statistics *38*, 197--214.

[Schwarz et al., 2009] Schwarz, K., Pizzini, A., Arendacka, B., Zerlauth, K., Filipiak, W., Schmid, A., Dzien, A., Neuner, S., Lechleitner, M., Scholl-Bürgi, S. et al. (2009). Breath acetone—aspects of normal physiology related to age and gender as determined in a PTR-MS study. Journal of Breath Research *3*, 027003.

[Searle et al., 1992] Searle, S., Casella, G. and McCulloch, C. (1992). Variance Components. Wiley Series in Probability and Statistics, Wiley.

[Simon et al., 2013] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. Journal of Computational and Graphical Statistics *22*, 231--245.

[Tan et al., 2018] Tan, Z., Roche, K., Zhou, X. and Mukherjee, S. (2018). Scalable Algorithms for Learning High-Dimensional Linear Mixed Models. arXiv preprint arXiv:1803.04431 .

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) *58*, 267--288.

[Tibshirani, 2014] Tibshirani, R. J. (2014). Degrees of Freedom and Model Search. arXiv preprint arXiv:1402.1920.

[Tibshirani et al., 2013] Tibshirani, R. J. et al. (2013). The lasso problem and uniqueness. Electronic Journal of Statistics *7*, 1456--1490.

[Vaida and Blanchard, 2005] Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. Biometrika *92*, 351--370.

[Waldron et al., 2011] Waldron, L., Pintilie, M., Tsao, M.-S., Shepherd, F. A., Huttenhower, C. and Jurisica, I. (2011). Optimized application of penalized regression methods to diverse genomic data. Bioinformatics *27*, 3399--3406.

[West et al., 2006] West, B., Welch, K. B. and Galecki, A. T. (2006). Linear mixed models: a practical guide using statistical software. CRC Press.

[Wray et al., 2010] Wray, N. R., Yang, J., Goddard, M. E. and Visscher, P. M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. PLoS genetics *6*, e1000864.

[Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *68*, 49--67.

[Zhao and Yu, 2006] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. The Journal of Machine Learning Research *7*, 2541--2563.

[Zhou et al., 2013] Zhou, X., Carbonetto, P. and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. PLoS genetics *9*, e1003264.

[Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association *101*, 1418--1429.

Práca autora

Články uverejnené v časopisoch

- Coufal, D., Jakubík, J., Jajcay, N., Hlinka, J., Krakovská, A., & Paluš, M. (2017). Detection of coupling delay: A problem not yet solved. Chaos: An Interdisciplinary Journal of Nonlinear Science, 27(8), 083109.
- Krakovská, A., Jakubík, J., Chvosteková, M., Coufal, D., Jajcay, N., & Paluš, M. (2018). Comparison of six methods for the detection of causality in a bivariate time series. Physical Review E, 97(4), 042207.

Iné publikované články

- Krakovská, A., Jakubík, J., Budáčová, H., & Holecyová, M. (2015). Causality studied in reconstructed state space. Examples of uni-directionally connected chaotic systems. arXiv, 1511.00505 , v1.

Články zaslané do časopisu

- Jakubík, J. (2018). Convex method for selection of fixed effects in high-dimensional linear mixed models. arXiv, 1805.09700, v1.

Príspevky uverejnené v zborníkoch konferencií

- Jakubík, J., *Convex variable selection for high-dimensional linear mixed models*., CFE-CMStatistics 2017 Book of Abstracts, 16-18 December 2017, London, UK, s 125. ISBN 978-9963-2227-4-2.
- Jakubík, J., *Comparison of methods for selection of fixed variables in high-dimensional linear mixed models*, Proceedings of the 19th European Young Statisticians Meeting, August 31 – September 4, 2015 Prague, s. 64-68. Praha, Czech Republic, ISBN 978-80-7378-301-3.
- Jakubík, J., *Comparison of methods for variable selection in high-dimensional linear mixed models*, Proc. of the 10th Int. Conf. on Measurement, Smolenice, Slovakia, May 25-28, 2015, s.55-58. Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, ISBN 978–80–969672–9–2.
- Jakubík, J., *LINEAR MIXED MODELS FOR GENOME-WIDE ASSOCIATION STUDIES*, Proceedings of the 6th International Young Biomedical Engineers and Researchers Conference, July 2-4, 2014. s. 160-163. Bratislava, Slovak Republic. ISBN 978–80–971697–0–1.

Príspevky prednesené na konferenciách

- ROBUST 2014, 18. zimní škola JČMF, ČR, 19.-24. január 2014
  Príspevok: Porovnanie funkcií na selekciu fixných efektov vo vysoko dimenzionálnych lineárnych zmiešaných modeloch.

- YBERC 2014, 6th International Young Biomedical Engineers and Researchers Conference, July 2-4, 2014. Bratislava, Slovak Republic
  Príspevok: LINEAR MIXED MODELS FOR GENOME-WIDE ASSOCIATION STUDIES.

- MEASUREMENT 2015, 10th Int. Conf. on Measurement, Smolenice, Slovakia, May 25-28, 2015.
  Príspevok: Comparison of methods for variable selection in high-dimensional linear mixed models

- PROBASTAT 2015 - 7th International Conference on Probability and Statistics. June 29 – July 3, 2015. Smolenice Castle, Slovakia.
  CONVEX METHOD FOR VARIABLE SELECTION IN HIGH-DIMENSIONAL LINEAR MIXED MODELS

- 19th European Young Statisticians Meeting. Prague, The Czech Republic. 31 Aug - 4 Sep 2015.
  Comparison of methods for selection of fixed variables in high-dimensional linear mixed models

- ROBUST 2016, 19. letní škola JČMF, ČR, 11.-16. september 2016
  Príspevok: Výber regresorov v lineárnych zmiešaných modeloch s malým počtom prediktorov.

- CMStatistics 2017 - The 10th International Conference of the ERCIM WG on Computational and Methodological Statistics, London, UK, 16-18 December 2017. Príspevok: Convex variable selection for high-dimensional linear mixed models.

- ROBUST 2018, 20. zimní škola JČMF, ČR, 21.-26. január 2018
  Príspevok: Nelineárna Grangerova kauzalita pomocou neurónových sietí.