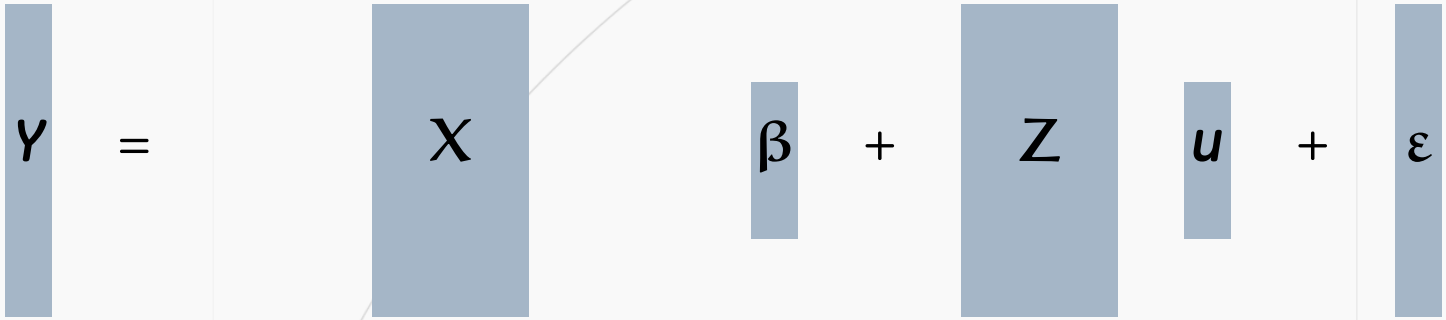
Výber regresorov v lineárnych zmiešaných modeloch s malým počtom prediktorov

Jozef Jakubík

Ústav merania
Slovenská akadémia vied

13. september 2016

Vysoko dimenzionálny lineárny zmiešaný model (LMM)

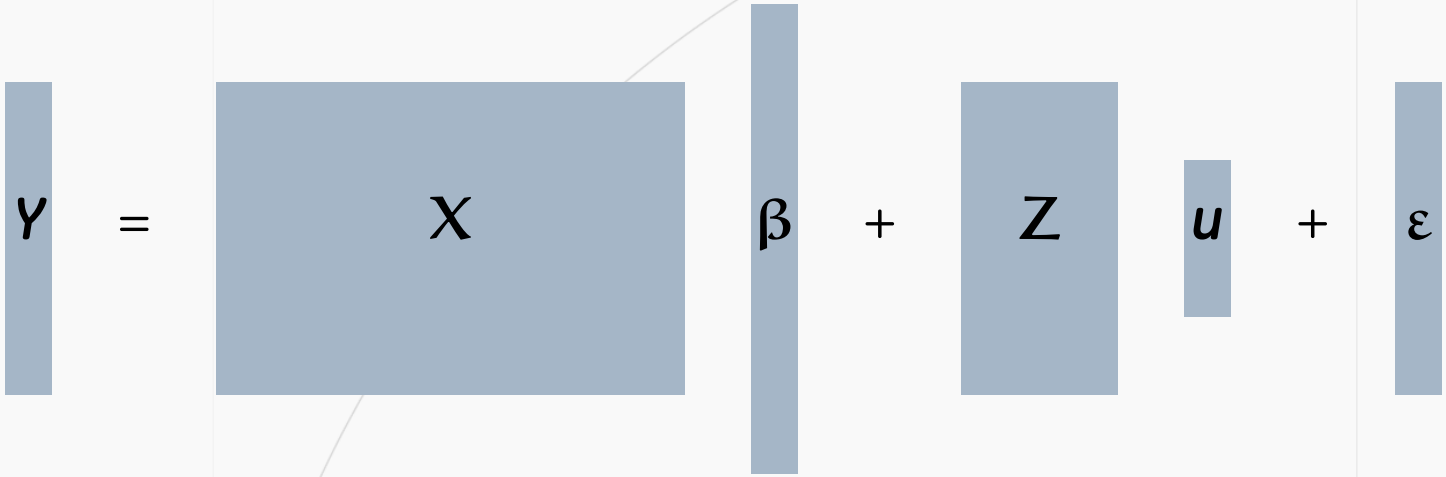


kde

- $\boldsymbol{Y}$ je $n \times 1$ vektor pozorovaní,
- $\mathbf{X}$ je $n \times p$ matica regresorov,
- β je $p \times 1$ vektor neznámych pevných efektov,
- $\mathbf{Z}$ je $n \times q$ matica prediktorov,
- $\boldsymbol{u}$ je $q \times 1$ vektor náhodných efektov z rozdelenia $\mathcal{N}(0, \mathbf{D})$,
- ε je $n \times 1$ vektor chýb z rozdelenia $\mathcal{N}(0, \mathbf{R} = \sigma_\varepsilon^2 \mathbf{I})$ a je nezávislý od $\boldsymbol{u}$.

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

Vysoko dimenzionálny lineárny zmiešaný model (LMM)



$$p >> n >> q$$

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

Výber premenných

Predpoklad:

- len malá skupina všetkých $\mathfrak{p}$ regresorov (z matice $\mathbf{X}$) ovplyvňuje pozorovania $\boldsymbol{Y}$ (len malá množina skutočných efektov $\boldsymbol{\beta}^0$ je nenulová). Túto podmnožinu označme S^0 a nech $s^0 = |S^0|$;
- všetky $\mathfrak{q}$ prediktory (z matice $\mathbf{Z}$) ovplyvňuje pozorovania $\boldsymbol{Y}$, ale efekt niektorých prediktorov môže byť malý.

Našim cieľom je odhadnúť S^0.

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

Motivácia

LMM dovoľujú lepšie špecifikovať kovariančnú štruktúru modelov, čo umožňuje lepšie popísať vzťahy v dátach.

Napríklad:

- populačná štruktúra,
- rodinné vzťahy.

Toto môže byť užitočná napríklad v Genome-wide association studies (GWAS).

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

Existujúce metódy

- LMMLASSO
- LASSOP

ℓ_1 penalizované odhady odvodené metódou maximálnej vierohodnosti

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

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

Konvexná metóda - LMMconvexLASSO

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

Podmienka:

$$q + s_{\lambda_0, \Lambda} < n$$

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

Voľba váh

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

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

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

Vybratá množina regresorov

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

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

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

Znamienková konzistencia

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

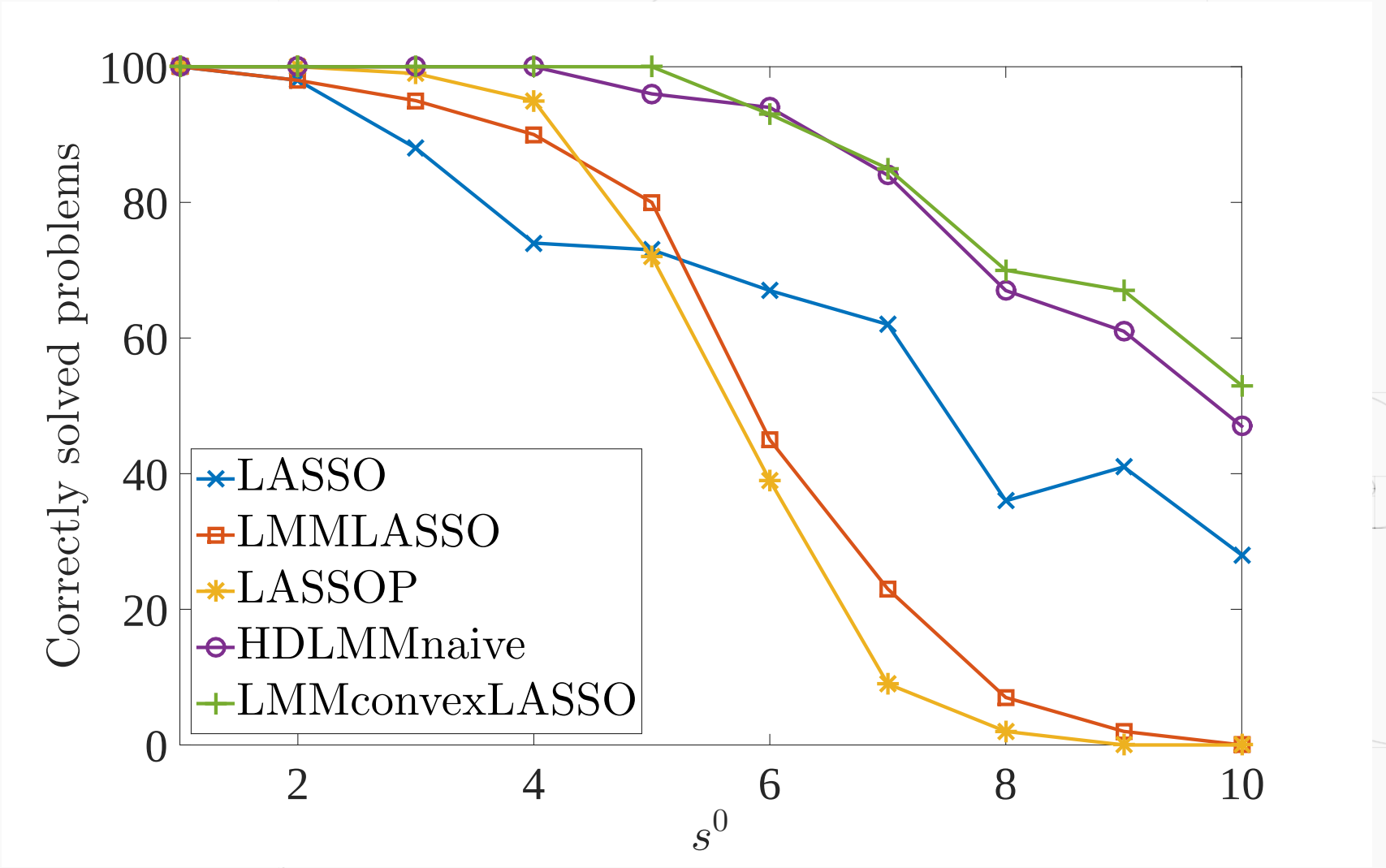
kde $\hat{\beta}^n(\lambda^n, \Lambda) =_s \beta^0$ znamená $\text{sign}(\hat{\beta}^n(\lambda^n, \Lambda)) = \text{sign}(\beta^0)$.

Navrhnutá 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 irrepresentable condition. Zároveň platí:

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

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

$n = 120, p = 150, q = 40$



Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

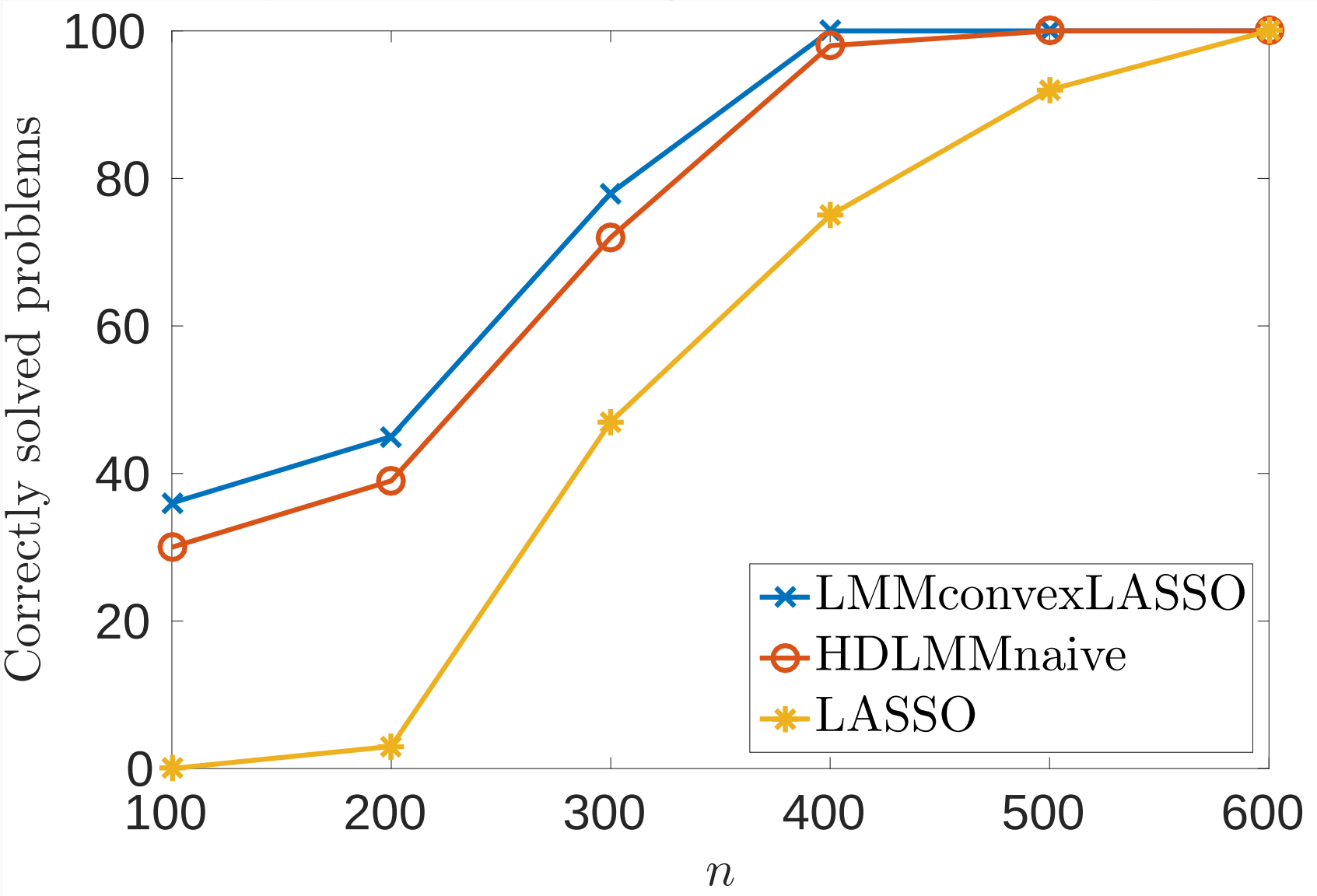
Simulačná štúdia

- $n = 200$
- $p = 5000$
- $q = 40$
- Matica $\mathbf{D}$ je diagonálna s dvoma varinčnými komponentami $\sigma_1^2 = 1.2$ a $\sigma_2^2 = 0.8$, $\sigma_\varepsilon^2 = 0.2$.
- Prvých $s^0 = 5$ 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.

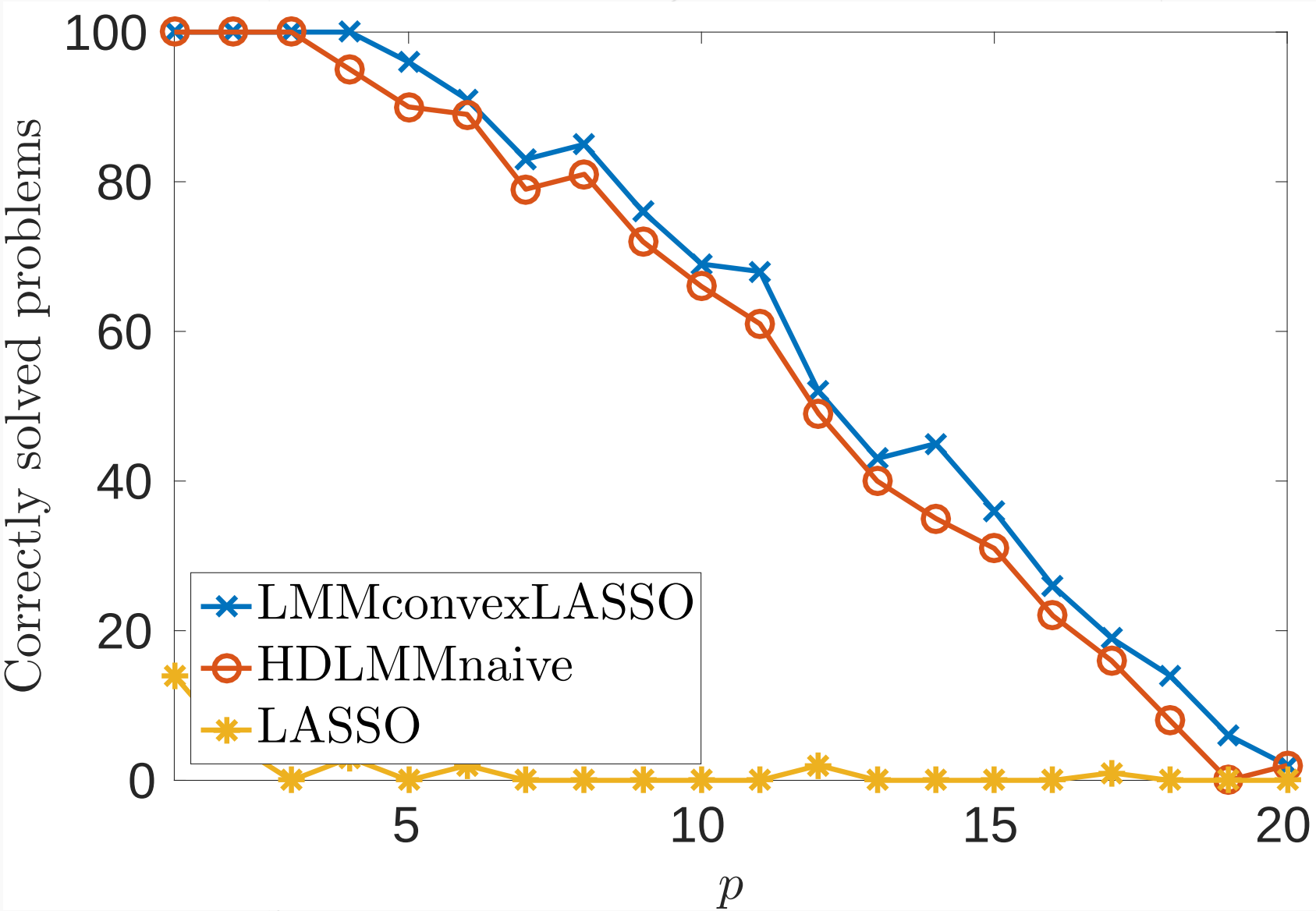
Pre každú metódu budeme sledovať v koľkých zo sto prípadov nám daná metóda poskytne aspoň pre jeden parameter (skupinu parametrov) práve množinu S^0.

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

n ($s^0 = 15$)

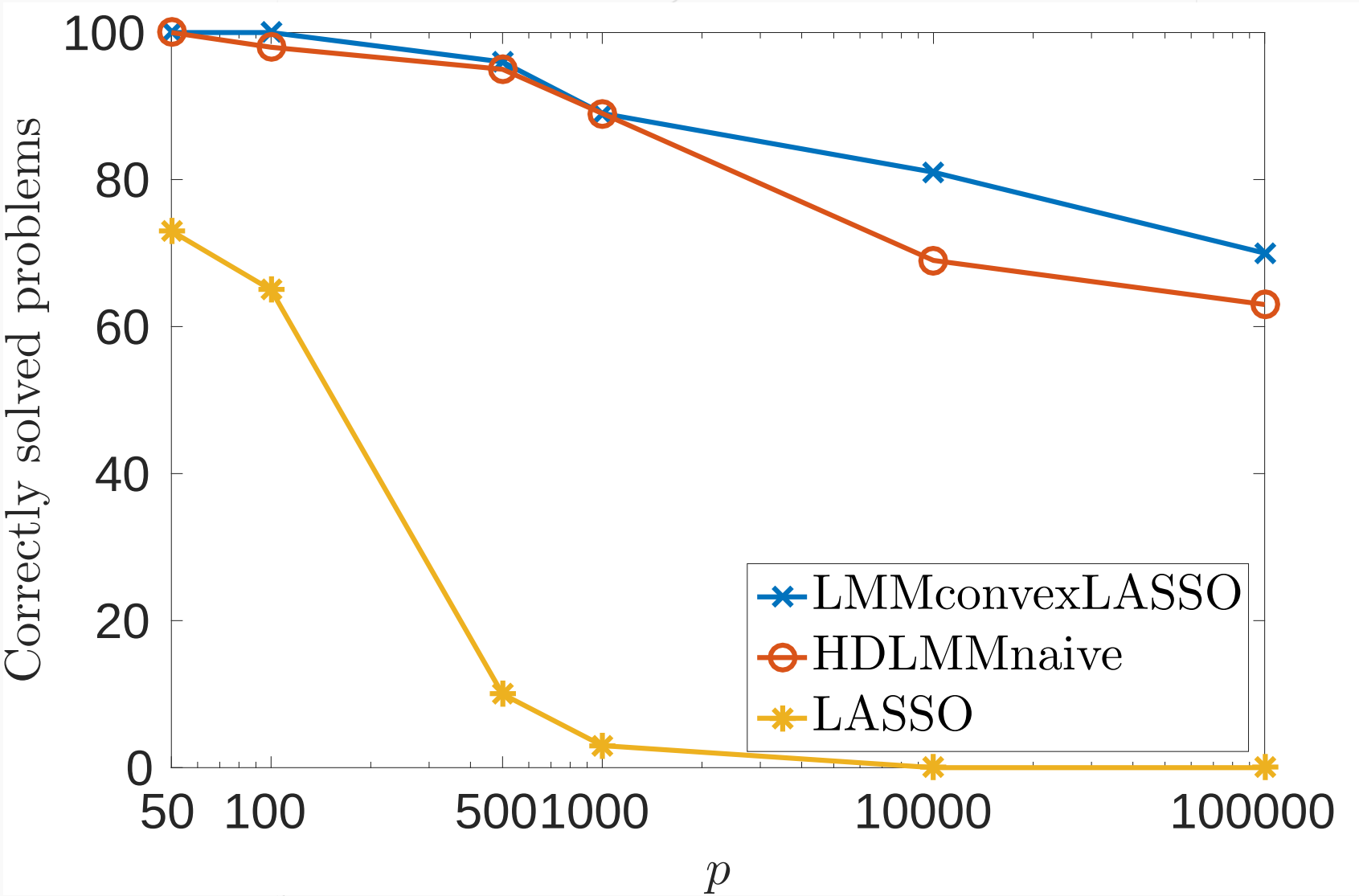


Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied



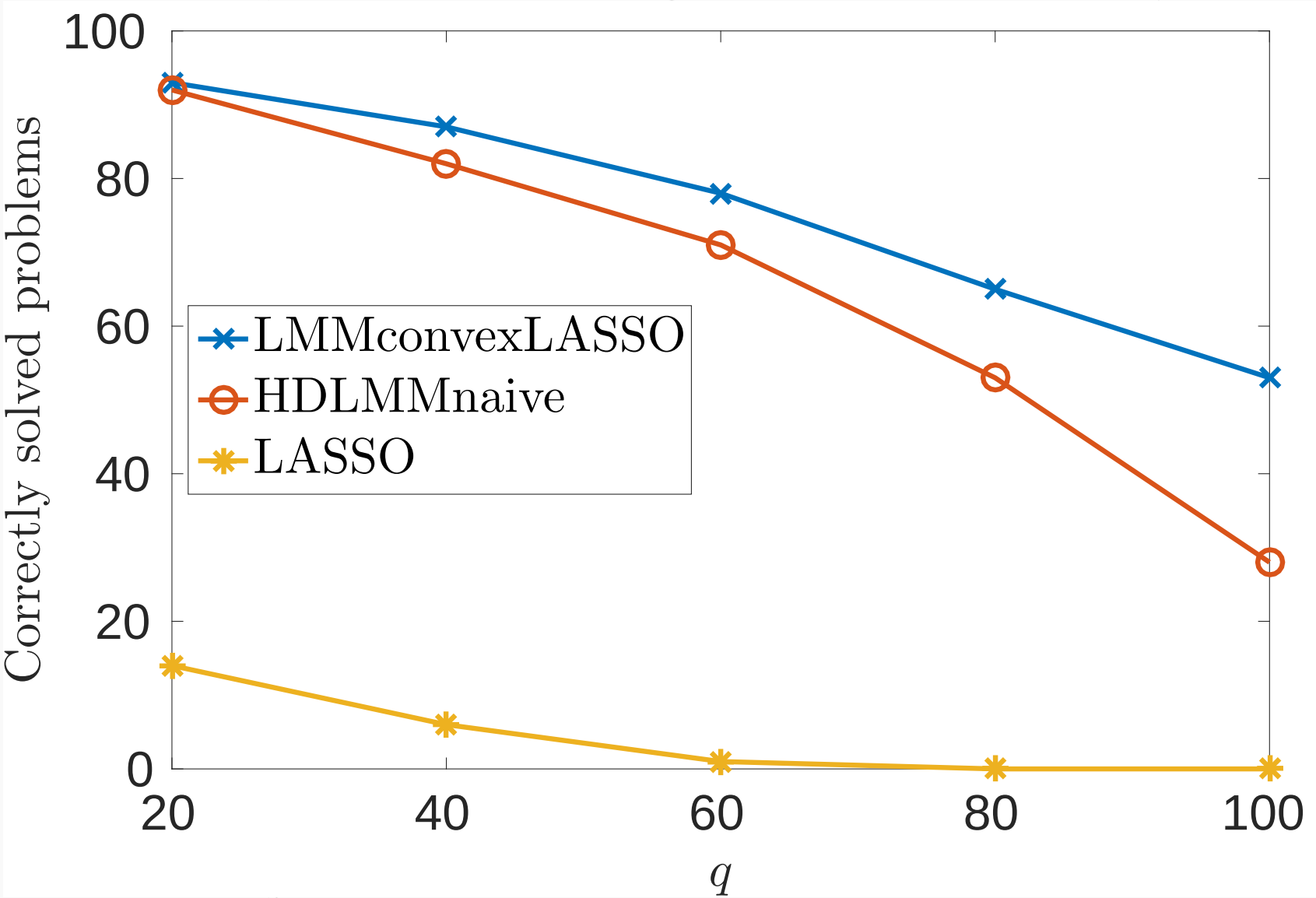
Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

p



Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

q



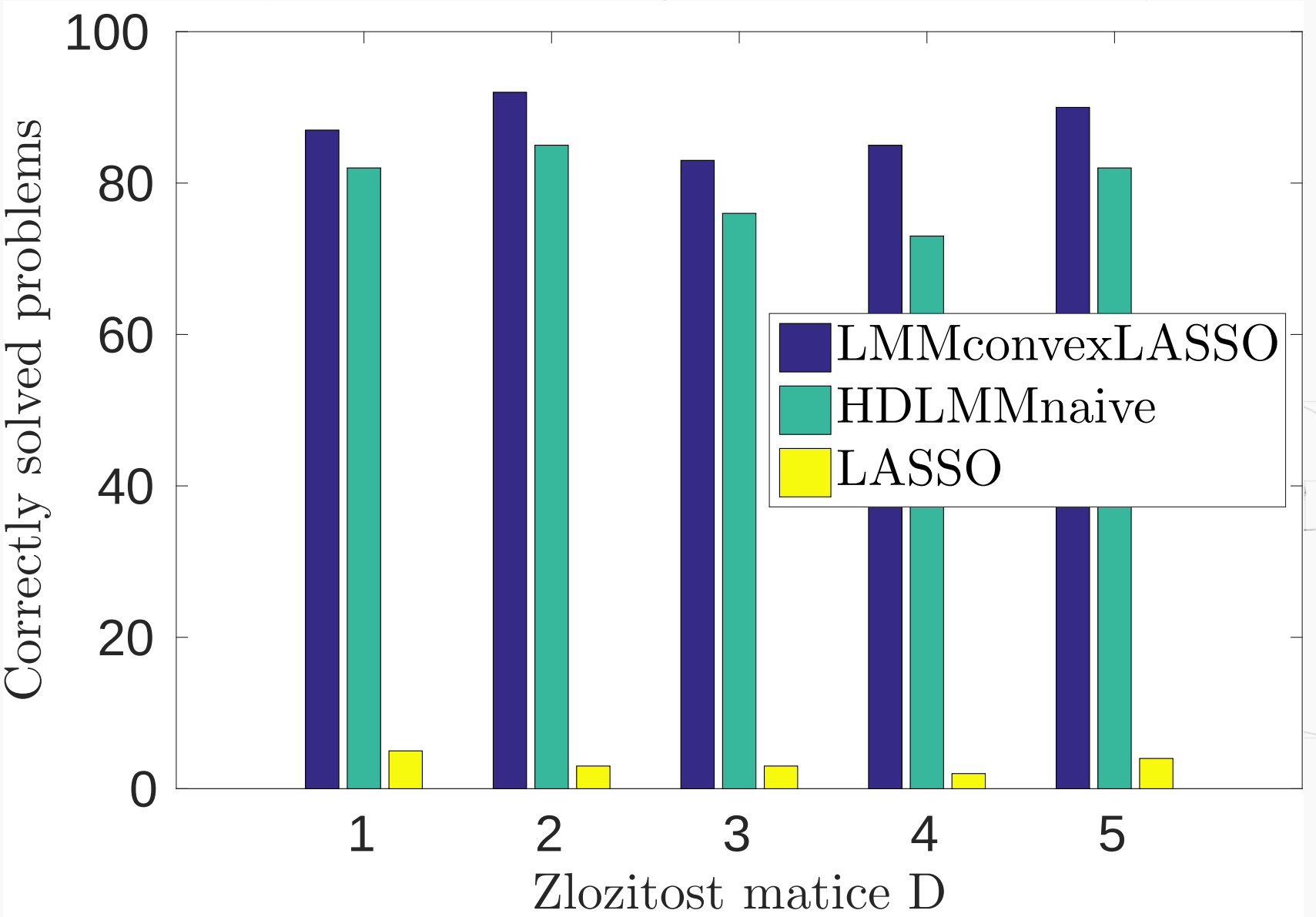
Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

D

- V prvom prípade bude kovariančná matica diagonálna s jednotkami na diagonále.
- V druhom prípade bude kovariančná matica diagonálna s dvoma časťami $\sigma_1^2 = 1.2$ a $\sigma_2^2 = 0.8$.
- V ďalšom prípade bude kovariančná matica blokovo diagonálna s diagonálou ako v predchádzajúcom prípade a prvky pod a nad hlavnou diagonálou budú mať hodnotu 0.2.
- V ďalšom prípade budú na prvých troch vedľajších diagonálach postupne hodnoty 0.3, 0.2 a 0.1.
- V poslednom prípade budú oba bloky obsahovať plné matice s hodnotou 0.2 mimo diagonály.

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

D



Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Ústav merania, Slovenská akadémia vied

Literatúra

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

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, 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.

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

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