Secondo incontro sul tema:

"Prospettive di ricerca allo IASI in Fisiopatologia, Bioinformatica e Biomatematica - Parte II"

Giovanni Felici

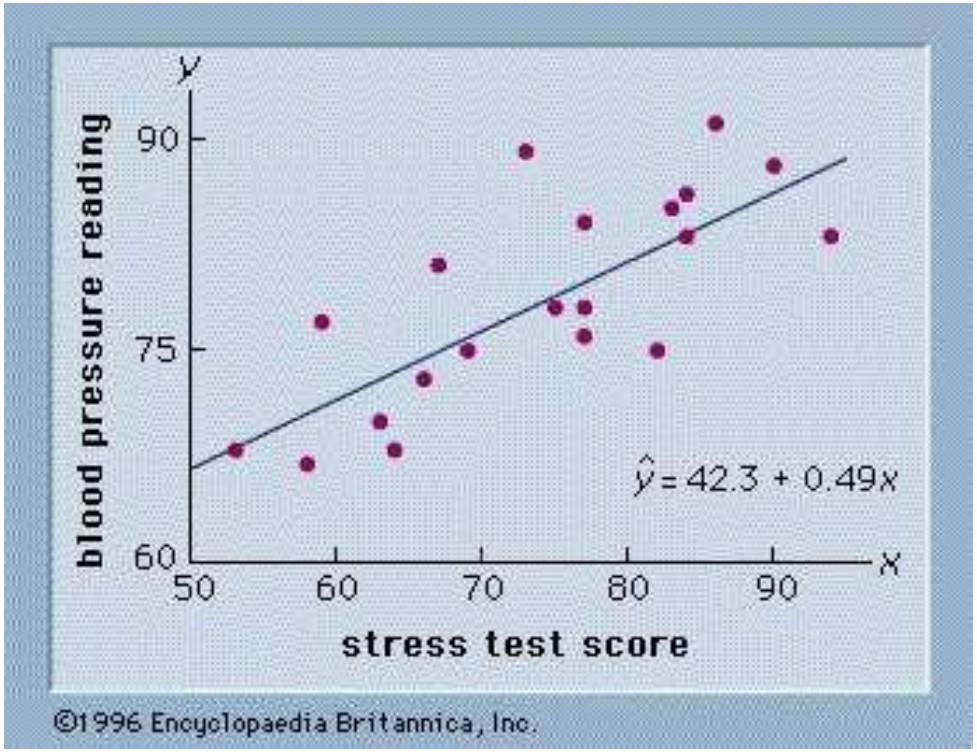
Zero-norm regularization of linear regression: could it be useful in the analysis of biological data ?

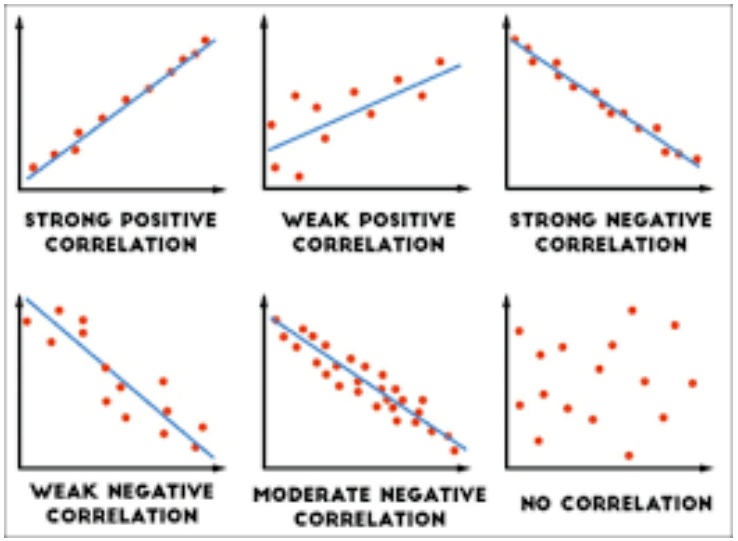
Main results based on paper:

MIP-BOOST: Efficient and Effective L_0 Feature Selection for Linear Regression,
A. Kenney, F.Chiaromonte, G. Felici, https://arxiv.org/abs/1808.02526
Journal of Computational and Graphical Statistics, 2020, to appear

Linear regression…. Simple, yet…

$$Y = b_0 + b_1 X_1 + e$$





$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_p X_p + e$$

Linear regression…. Simple, yet…

$$Loss(h_w) = \sum_{i=1}^{n} [y_i - (w_0 + w_1 x_i)]^2$$

$$\frac{\partial}{\partial w_0} \sum_{i=1}^{n} [y - (w_0 + w_1 x_1)]^2 = 0 \qquad w_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

$$\frac{\partial}{\partial w_1} \sum_{i=1}^{n} [y - (w_0 + w_1 x_1)]^2 = 0 \qquad w_1 = \frac{N \sum (x_i y_i) - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

- **p** variables, or features, **n** observations / samples

$$Y = X^t \beta + \varepsilon$$

$$\hat{\beta} = \operatorname{argmin}_{\beta \in R^p} \left\| Y - X^t \beta \right\|_2$$

$$\hat{\beta} = \left(X^t X \right)^{-1} X^t Y$$

To invert covar matrix, $n \geq p$

- What if **p** is large, **p >> n** ?

- **Sparsification / Feature Selection**: use few terms
- **Regularization**: force coefficients to be “nice” in size

In biological applications

- **p >> n ?** May happen…

- Gene expression data
- CNV data
- Clinical data
- Proteomics, omics in general

-> very often, the variables are many many, the samples are few few (they have indeed a large cost!)

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_pX_p + e$$

- Do we need all p of them for a “good” regression?
- Are some variables “correlated” (sort of, equal)
- Don’t we **overfit** with all these variables ?
- Is our model robust enough with all these variables ? Noise is always around the corner…

Regularization = sparisification?

$$\min_{\beta} \left\| Y - X^t \beta \right\|_q \quad s.t. \left\| \beta \right\|_r \leq s$$

$$\min_{\beta} \left\| Y - X^t \beta \right\|_q + \lambda \left\| \beta \right\|_r$$

Error Penalty

r = 2 → *Ridge Regression* (Tikhonov regularization)

r = 1 → *Lasso*

R = 1 & 2 → *Elastic Net*

- **Tibshirani, R**. (**1996**), "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society. Series B
- **Zou, H., and Hastie**, T. (**2005**), "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society: Series B.

Controlling the norm of β

Do RR or LASSO really ***sparsify*** the model ?

Convexity of feasible region for β → **Less dimension reduction**

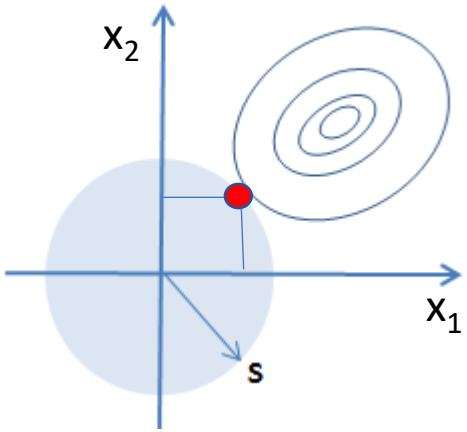
- Going from 2-norm to 1-norm cuts out portions of the feasible region where all coefficients have similar values
- the solution is pushed towards the vertices of the feasible region
- sparsification

$$\|\beta\|_r \le s$$

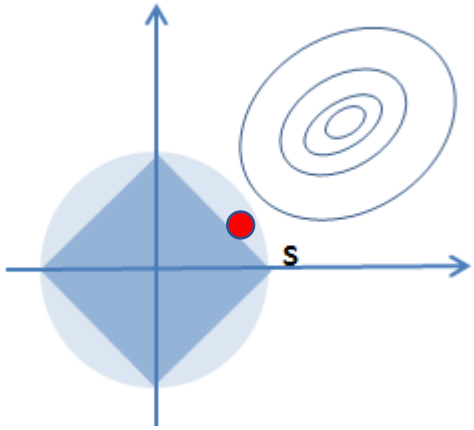
$$\sqrt{\sum_i \beta_i^{\,2}} \le 1$$

$$\left|\sum_i \beta_i\right| \le 1$$

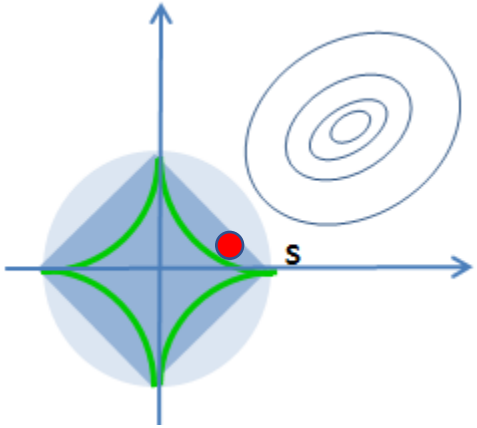
$$\left(\sum_i \beta_i^{\frac{1}{2}}\right)^2 \le 1$$



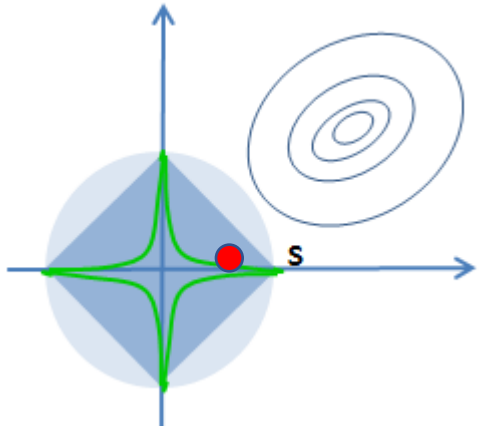
r = 2



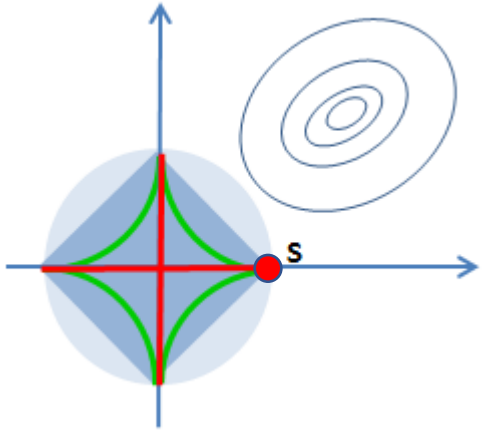
r = 1



r = r' < 1



r < r' < 1



r = 0

$\|\beta\|_0$ = *Number of nonzero elements*

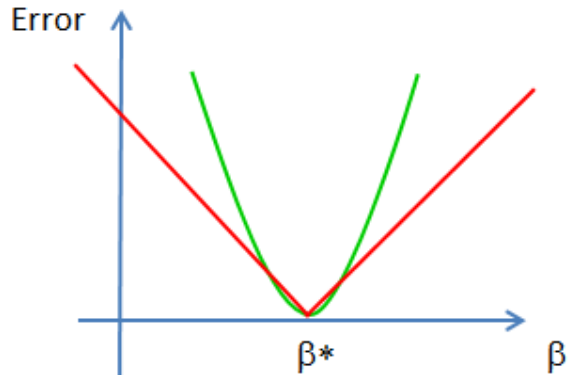
"0-norm" as a Mixed Integer Linear Problem

- Optimize under 0-norm constraints: **EASY** ? NO, exponential WC complexity
- **IMPOSSIBLE** ? NEITHER…
- Similar to subset selection: what is the subset of variables of given dimension k that minimizes my error?

Much easier using 1-**norm for the OF (error)**

Least Absolute Deviation (LAD), Robustness, etc…

$$\min \left\| Y - X^t \beta \right\|_2$$
$$s.t. \quad \left\| \beta \right\|_0 \leq s$$



$$\min \left\| Y - X^t \beta \right\|_1$$
$$s.t. \quad \left\| \beta \right\|_0 \leq s$$

“norm-0” as a MIQP

$$\min \sum_{i=1,n} e_i^2$$

$$s.t.$$

$$Y_i - \sum_{j=1,p} \beta_j X_{ij} \le e_i, \quad i = 1, n$$

$$Y_i - \sum_{j=1,p} \beta_j X_{ij} \ge -e_i, \quad i = 1, n$$

$$\beta_j \ge -Mz_j, \quad j = 1, p$$

$$\beta_j \le Mz_j, \quad j = 1, p$$

$$\beta_j \in R, z_j \in \{0,1\}, \quad j = 1, p$$

$$\sum_{j=1,p} z_j \le k$$

CONS

- Use binary variables to indicate if coefficients are used (1) or not (0)
- Numerically unstable constraints to link the choice and the value
- NP complexity class

PROS

- Extremely flexible
- Direct control of sparsity
- Many extensions
- Powerful optimization software to solve it efficiently
- Powerful heuristics

$$\min \sum_{i=1,n} e_i^2$$

$$s.t.$$

$$Y_i - \sum_{j=1,p} \beta_j X_{ij} \le e_i, \quad i = 1, n$$

$$Y_i - \sum_{j=1,p} \beta_j X_{ij} \ge -e_i, \quad i = 1, n$$

$$\beta_j \ge -Mz_j, \quad j = 1, p$$

$$\beta_j \le Mz_j, \quad j = 1, p$$

$$\beta_j \in R, z_j \in \{0,1\}, \quad j = 1, p$$

$$\sum_{j=1,p} z_j \le k$$

Not a new idea

- Bertsimas, D., & King, A. (**2017**). Logistic regression: From art to science. *Statistical Science*, *32*(3), 367-384.
- Bertsimas, D., & Van Parys, B. (**2017**). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *arXiv preprint arXiv:1709.10029*.
- Bertsimas, D., King, A., Mazumder, R. et al. (**2016**), "Best subset selection via a modern optimization lens," The Annals of Statistics, 44
- TSato, Takano, Miyashiro, Yoshise,(**2016**) Feature subset selection for logistic regression via mixed integer optimization, Computational Optimization and Applications 64 (3)
- Lan, Vucetic (**2013**), Multi-task feature selection in microarray data by binary integer programming, BMC proceedings Ichino, Sklansky (**1984**) Optimum feature selection by zero-one integer programming, IEEE Transactions on Systems, Man, and Cybernetics, 737-746

What we get for free:

$$\sum_{j=1,p} z_j \le k \quad \Rightarrow \quad Z \in S$$

$$\sum_{j \in G_1} z_j \le k_1, \quad \sum_{j \in G_2} z_j \le k_2$$

$$z_i + z_j \le 1$$

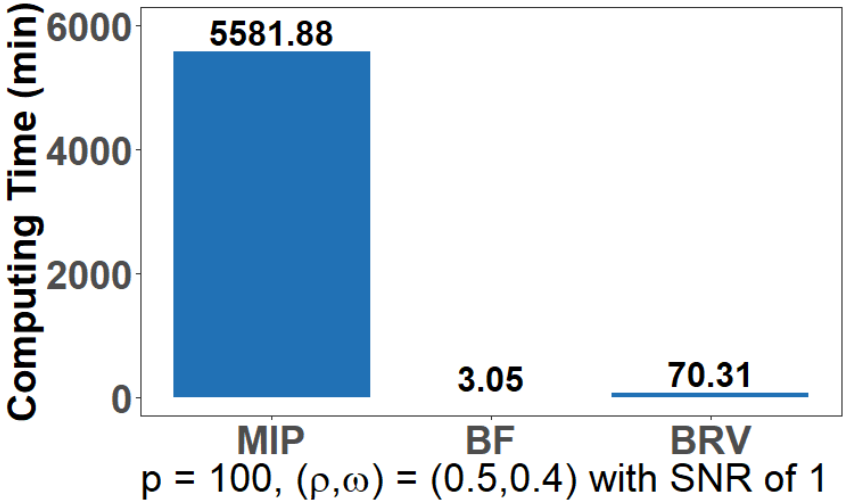
$$z_i \le z_j$$

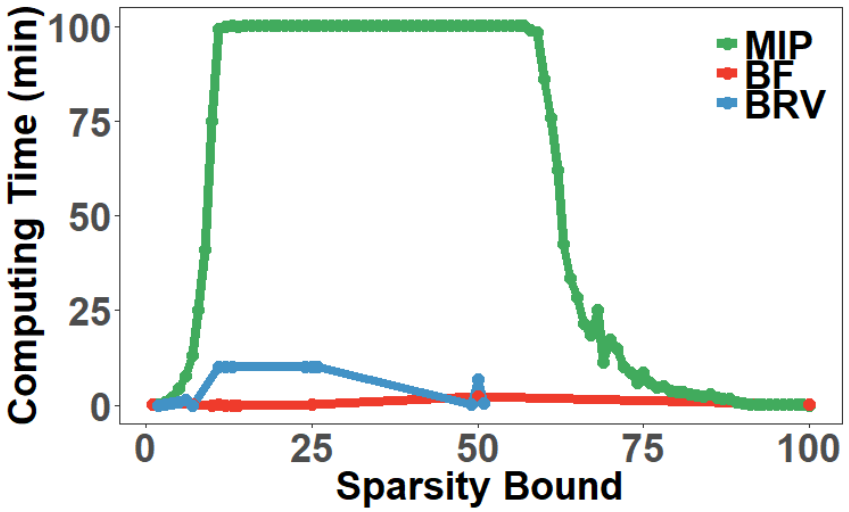
$$\sum_{j \in G} z_j \ge k'$$

$$\sum_{j \in G} c_j z_j \le B$$

Well, not for free: at computational cost

Would not it be nice?

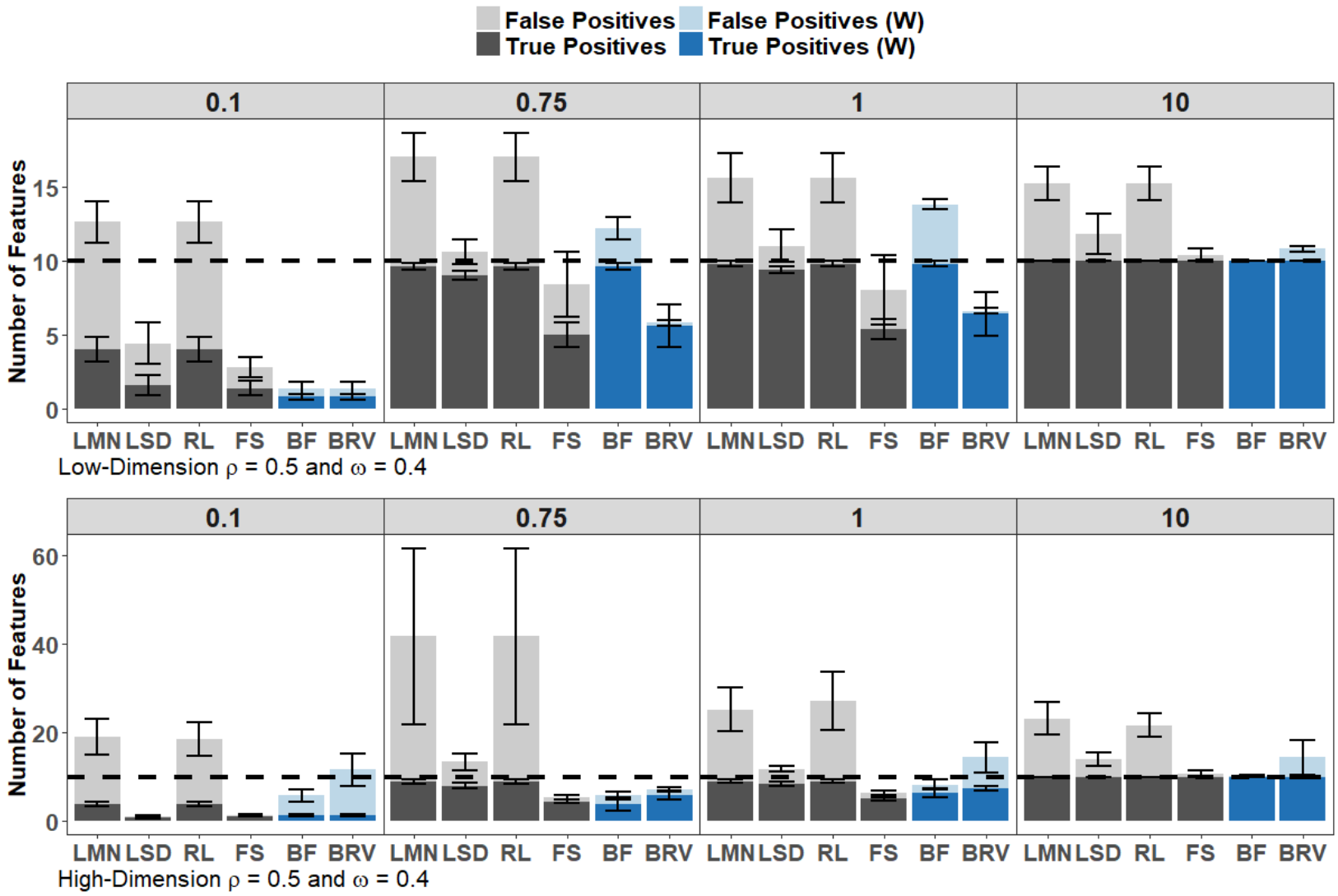




3 tricks

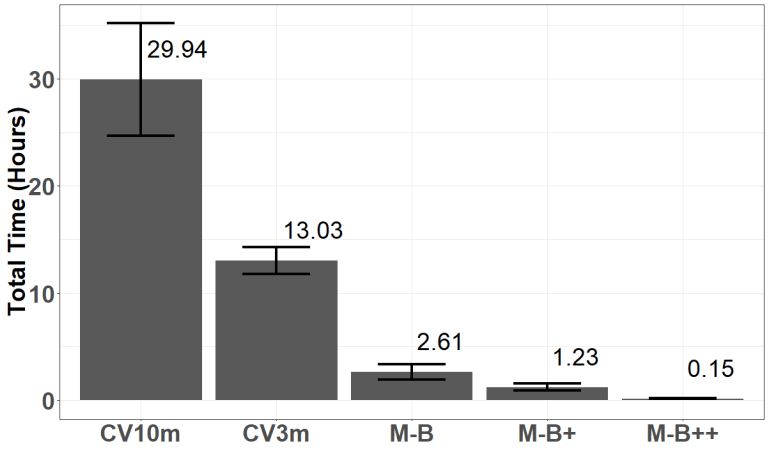
1. Bisection with feelers and randomly added variables
2. Integrated cross-validation
3. Whitening for feature selection

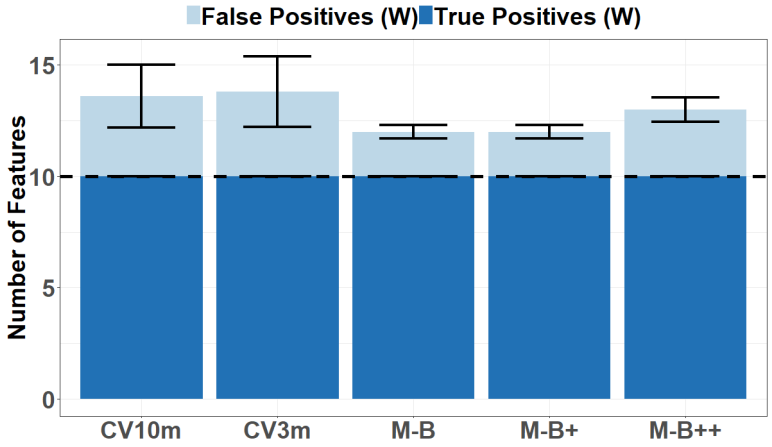
Sprser and Better solutions



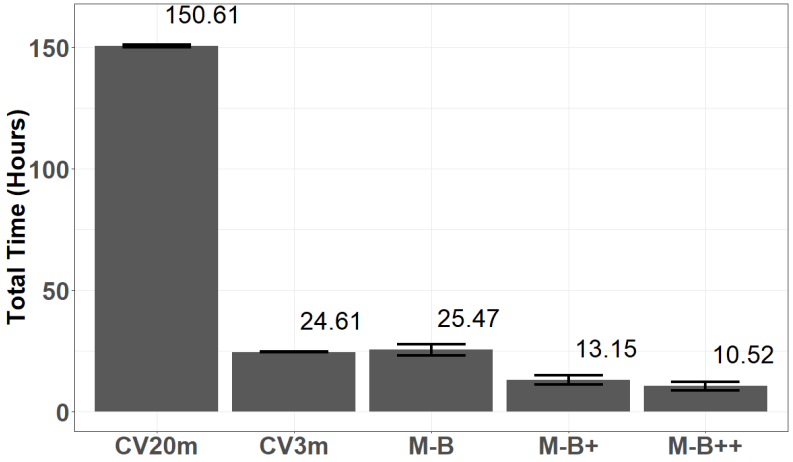
Time

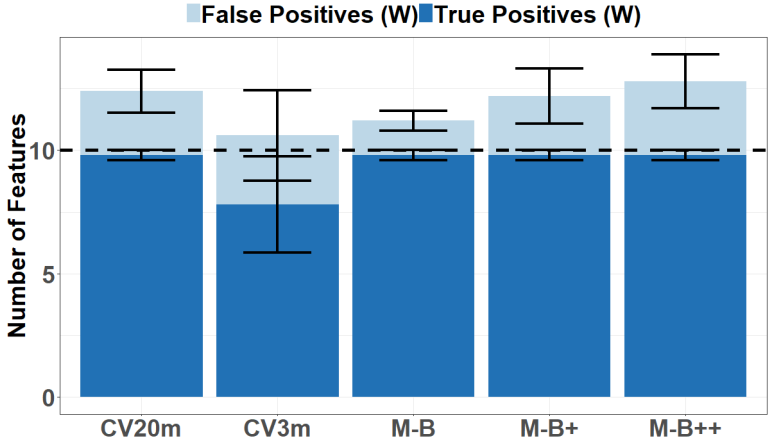
Low Dimension





High Dimension





Applications: Diabetes (Efron et al. (2004)), Bodyfat (Penrose et al. (1985))

Procedure	Diabetes		Bodyfat		Bodyfat with Density	
	$\hat{k}_0$	VALMSE	$\hat{k}_0$	VALMSE	$\hat{k}_0$	VALMSE
LMN	19	0.488	15	16.052	6	0.696
LSD	9	0.501	4	20.141	2	2.172
RL	19	0.488	15	16.052	6	0.696
FS	7	***0.477**	2	15.857	2	0.488
MIP-BOOST	3	0.496	13	***14.838**	1	***0.388**

*: minimum Validation MSE.

- LMN lasso with min.
- LSD lasso with parsim.
- RL relaxed lasso
- FS forward selection
- 25% validation

Diabetes

- n = 442 diabetes patients
- 10 baseline predictors (age, body mass index, average blood pressure, six blood serum measurements and sex)
- p = 64 features including 45 pair-wise products and 9 squared terms
- Y = disease progression

Bodyfat

- n = 252 patients
- 13 predictors (age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and writs circumference)
- Y = % of body fat (based on density…)
- inclusion of inbteractions on poly terms
- p = 559 / 679

- Lamotrigine (LTG)
- Body Mass Index
- Mean Arterial Pressure

- Without density: some combination of age, weight, height, abdomen, and wrist circumference

Thank you