Living in a Sparse World

Ernst Wit
Johann Bernoulli Institute
University of Groningen

e.c.wit@rug.nl
http://www.math.rug.nl/∼ernst

12 April 2010

Joint work with Luigi Augugliaro (Palermo)

rijksuniversiteit groningen

Numquam ponenda est pluralitas sine necessitate

William Occam (1288-1348) proposed a meta-theory of knowledge:

> *"For nothing ought to be posited without necessity."*

Can be interpreted as a

- **Aesthetic principle:** enhances model interpretability through parsimonious representation
- **Pragmatic principle:** computability.
- **A priori information principle:** represents expectation about nature of solution.
- **Prediction principle:** bias-variance trade-off
- **Bayesian principle?**

Exercise: Predict the next two numbers in the sequence

$$-1, 3, 7, 11, \ldots$$

Occam's razor does not have to be an aesthetic principle...

You probably thought $15, 19$, using the hypothesis

H_1 : *Add 4 to the previous number.*

But why not $-19.9, 1043.8$, using the hypothesis

H_2 : *if x is current number, then next number is given by*

$$-x^3/11 + 9/11x^2 + 23/11.$$

Relative evidence of two hypotheses given data D:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(H_1)}{P(H_2)} \frac{P(D|H_1)}{P(D|H_2)}$$

where $P(H_1)/P(H_2)$ could include an "aesthetic Occam" preference.

rijksuniversiteit groningen

... let the data speak for themselves!

Simple hypothesis uses two parameters:
(i) α=starting value and (ii) β=increment, assumed randomly drawn between -50 and $+50$, in which case

$$P(D|H_1) = P(\alpha = -1; \beta = 4) = \left(\frac{1}{101}\right)^2 = 1 \times 10^{-4}.$$

Complex hypothesis uses (arguably) 7 parameters:
(i) α=starting value and (ii-vii) $\frac{\beta_1}{\beta_2}$, $\frac{\beta_3}{\beta_4}$, $\frac{\beta_5}{\beta_6}$ in the polynomial, again, assumed randomly drawn between -50 and $+50$, in which case

$$P(D|H_2) = \ldots = \frac{1}{101}\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{2}{101}\frac{1}{50}\right) = 2.5 \times 10^{-12}.$$

So, $P(H_1|D)/P(H_2|D) = 4 \times 10^7$ IN FAVOUR OF SPARSITY.

rijksuniversiteit groningen

Example I: deletion/amplification of DNA

It is known that deletion and amplification of certain parts of DNA plays a role in the severity of breast-cancer.

John Bartlett (Royal Infirmary, Glasgow) wants to use deletion and amplification data on **62** breast cancer patients across **59** genes.

His expectation: a few gene deletions and gene amplifications will affect the severity of the breast cancer (measured as NPI).

$$\text{NPI}_i = \sum_{j=1}^{59} x_{ij}\beta_j, + \epsilon_i \qquad (\text{patient} i = 1, \ldots, 62),$$

subject to sparsity, i.e. many $\beta_j \approx 0$.

rijksuniversiteit groningen

Penalized inference: operational definition of sparsity

High-dimensional inference. As we want lots of small β, we consider the constraint maximization of

$$l(\beta) \text{ subject to } ||\beta||_q \leq c,$$

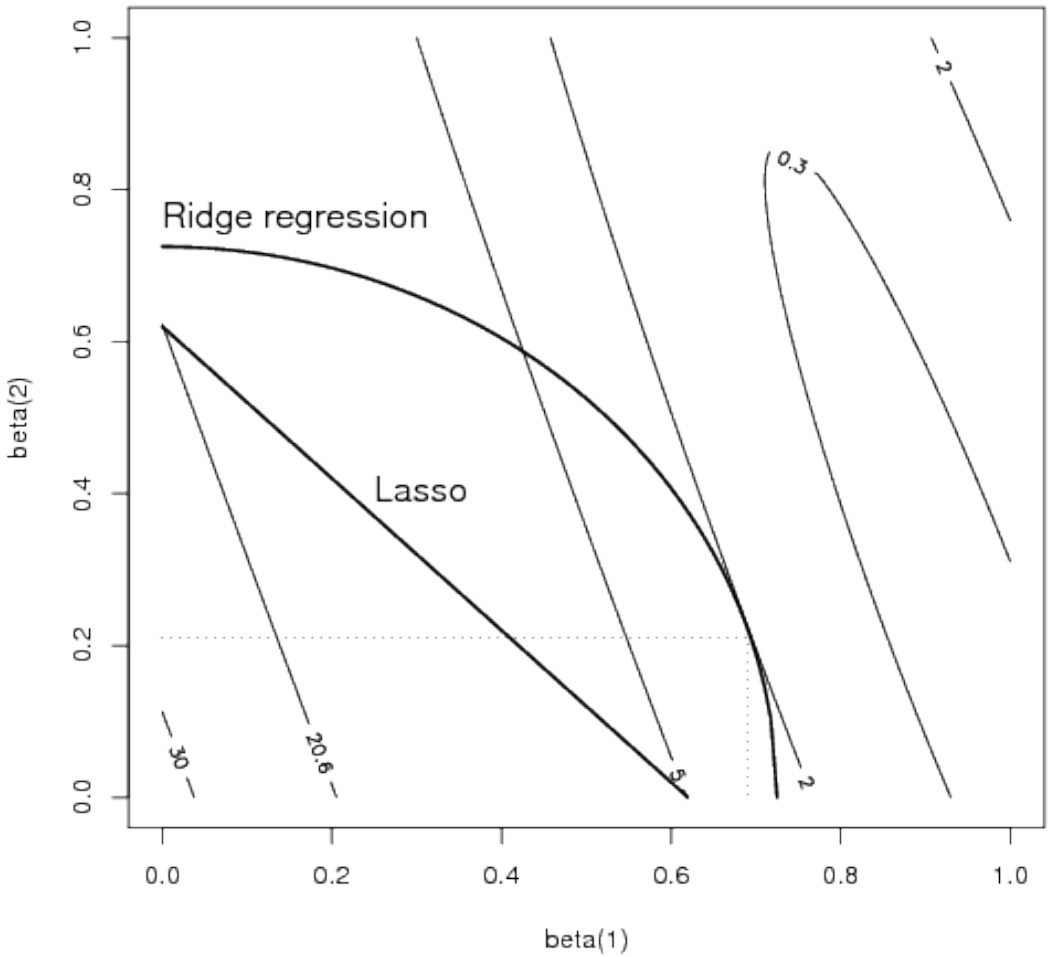
whose dual is equal to the penalized likelihood

$$l_\lambda(\beta) = l(\beta) - \lambda||\beta||_q^q$$

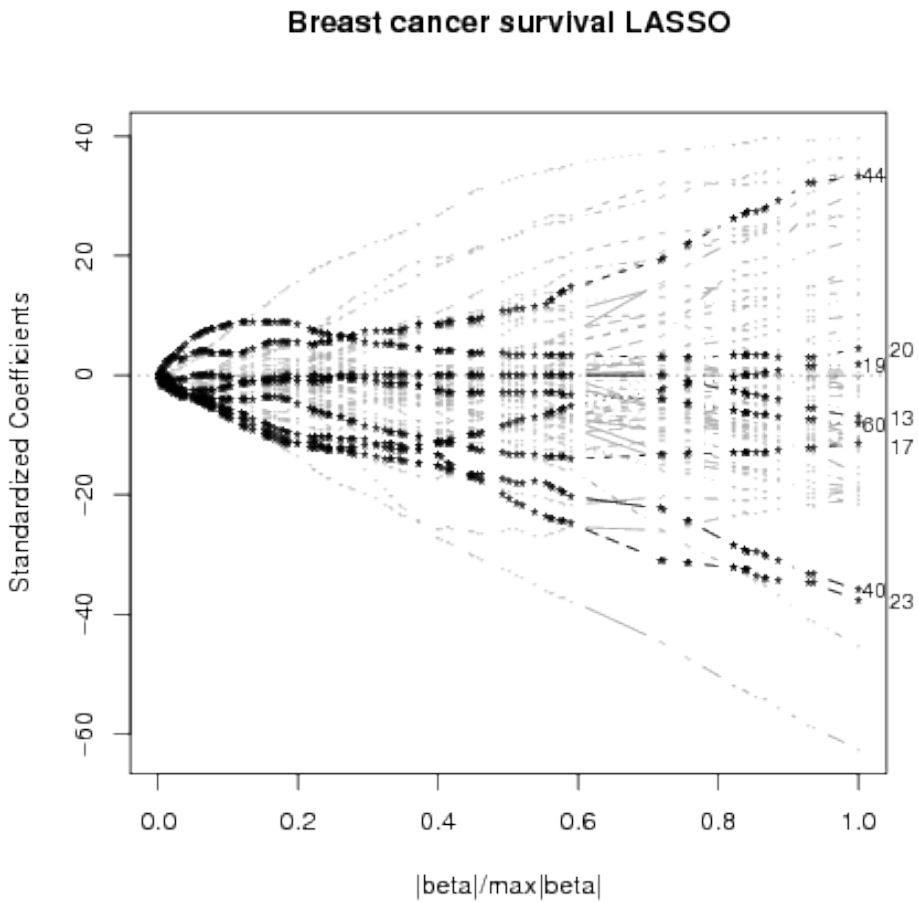
There are several special cases:

- $q = 2$: Ridge regression (1958)
  Under normality leads again to a simple quadratic form.
- $q = 1$: Lasso regression (Tibshirani, 1996)
  No closed form solution.

Geometry of the L_1 penalty = Sparsity

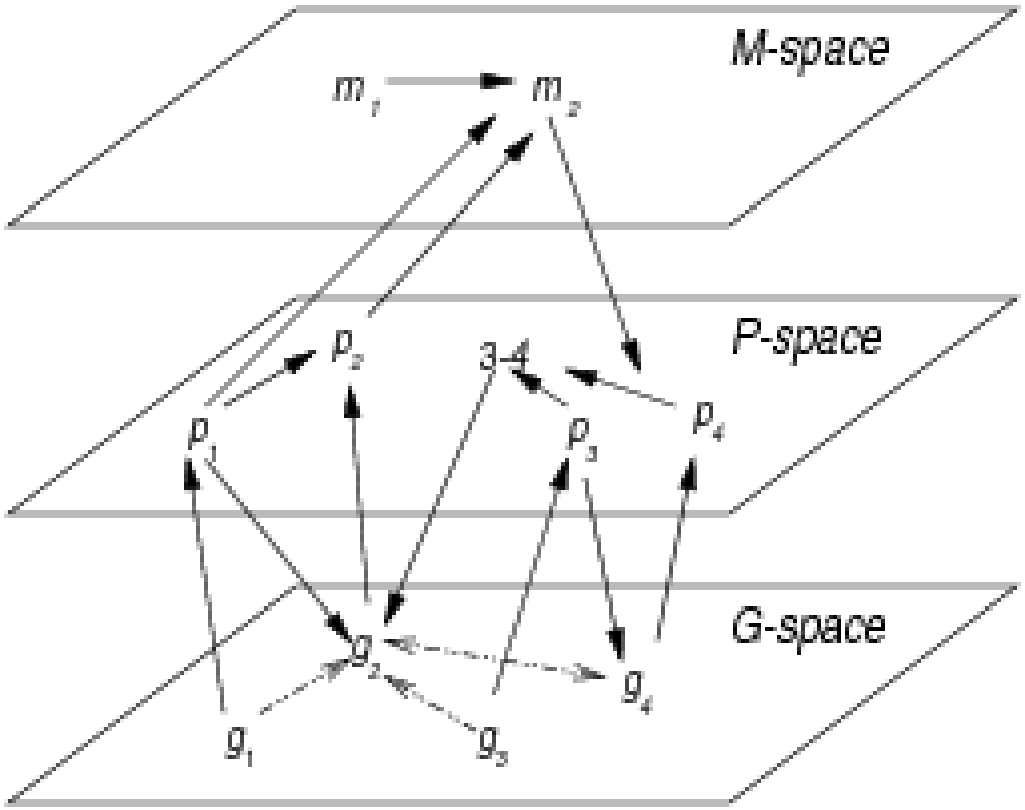


Lasso applied to DNA deletion/amplification data



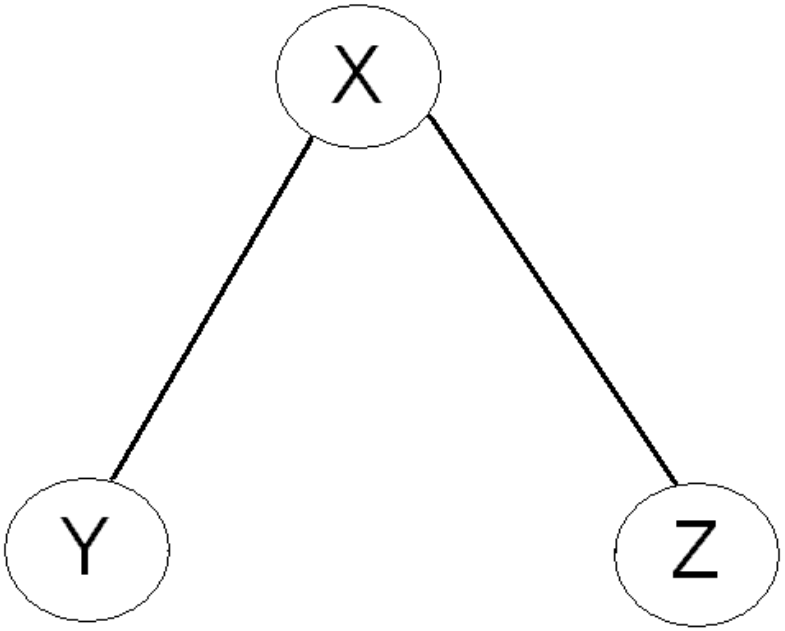
Some stopping rule selects 7 out of 59 genes.

Example II. gene regulatory networks



"A collection of DNA segments in a cell which "interact" with each other via their RNA or proteins and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed."

Gaussian Graphical Models



- Absence/presence of an edge indicates conditional independence of the variables:
- No edge joining Y and Z $\Longleftrightarrow$ $Y \perp Z | \text{rest}$

If $U = (X, Y, Z)$ in a graph are Gaussian, then

$$U \sim N(\mu, \Sigma).$$

whereby $\Theta = \Sigma^{-1}$ represents conditional independence, i.e.

if $\theta_{Y,Z} = 0$, then Y is independent of Z given rest.

rijksuniversiteit groningen

Estimating Θ

The Gaussian profile likelihood is given by

$$l(\Sigma) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y-\hat{\mu})^t\Sigma^{-1}(y-\hat{\mu}),$$

and so

$$\begin{aligned} l(\Theta) &= \frac{1}{2}\log|\Theta| - \frac{1}{2}(y-\hat{\mu})^t\Theta(y-\hat{\mu}) \\ &= \frac{1}{2}\left(\log|\Theta| - \text{Trace}(S\Theta)\right), \end{aligned}$$

where S the empirical covariance matrix of X. Then

$$\frac{\delta l(\Theta)}{\delta\Theta} = \Theta^{-1} - S,$$

and so $\widehat{\Theta} = S^{-1}$.

rijksuniversiteit groningen

Estimating the graph using the L_1 penalty

Use the *lasso* regularized log-likelihood:

$$\max_{\Theta} \left[\log |\Theta| - \text{Trace}(S\Theta) - \lambda ||\Theta||_1 \right],$$

with score equations:

$$\Theta^{-1} - S - \lambda \text{Sign}(\Theta) = 0.$$

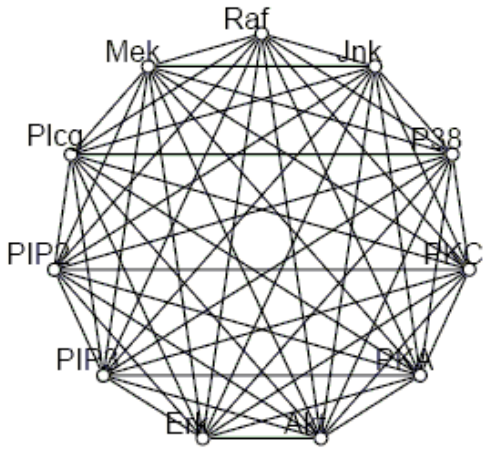
NOTE: Compare this with the Lasso problem:

$$\min_{\beta} (y - X\beta)^t (y - X\beta) + \lambda ||\beta||_1$$

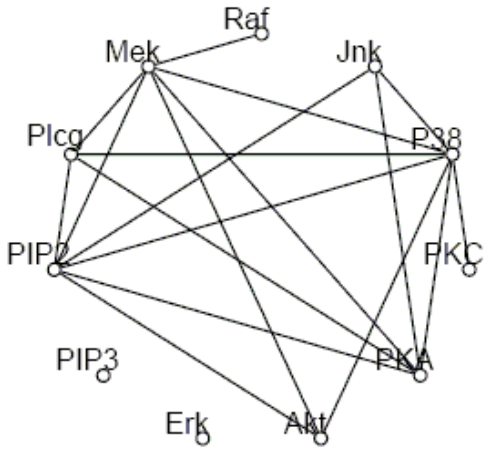
with solution

$$X^t X \beta - X^t y + \lambda \text{Sign}(\beta) = 0.$$

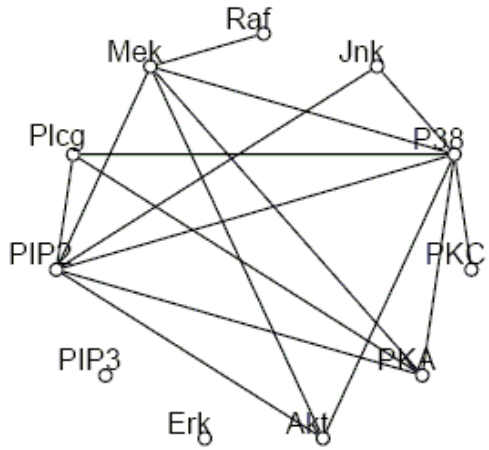
An example: Cell signalling network for different λ



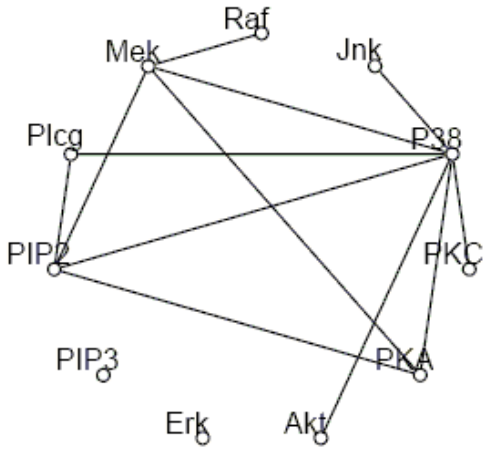
L1 norm= 2.27182



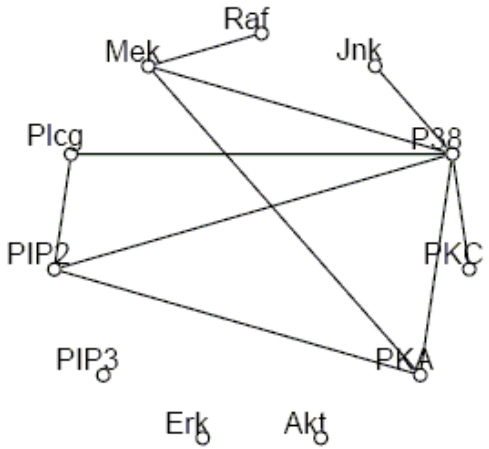
L1 norm= 0.08915



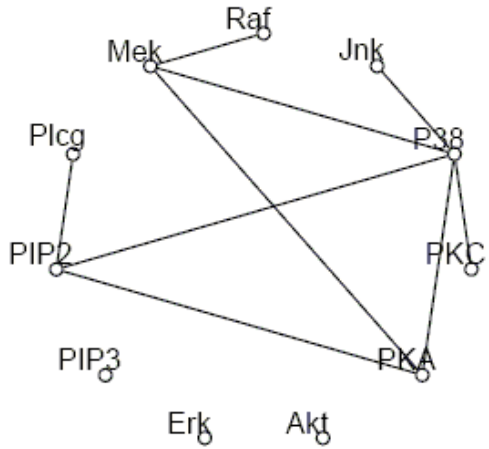
L1 norm= 0.04251



L1 norm= 0.02171



L1 norm= 0.01611



L1 norm= 0.01224

rijksuniversiteit groningen

III. Relationship L_1 and angles between x variables

Let's return to our old problem:

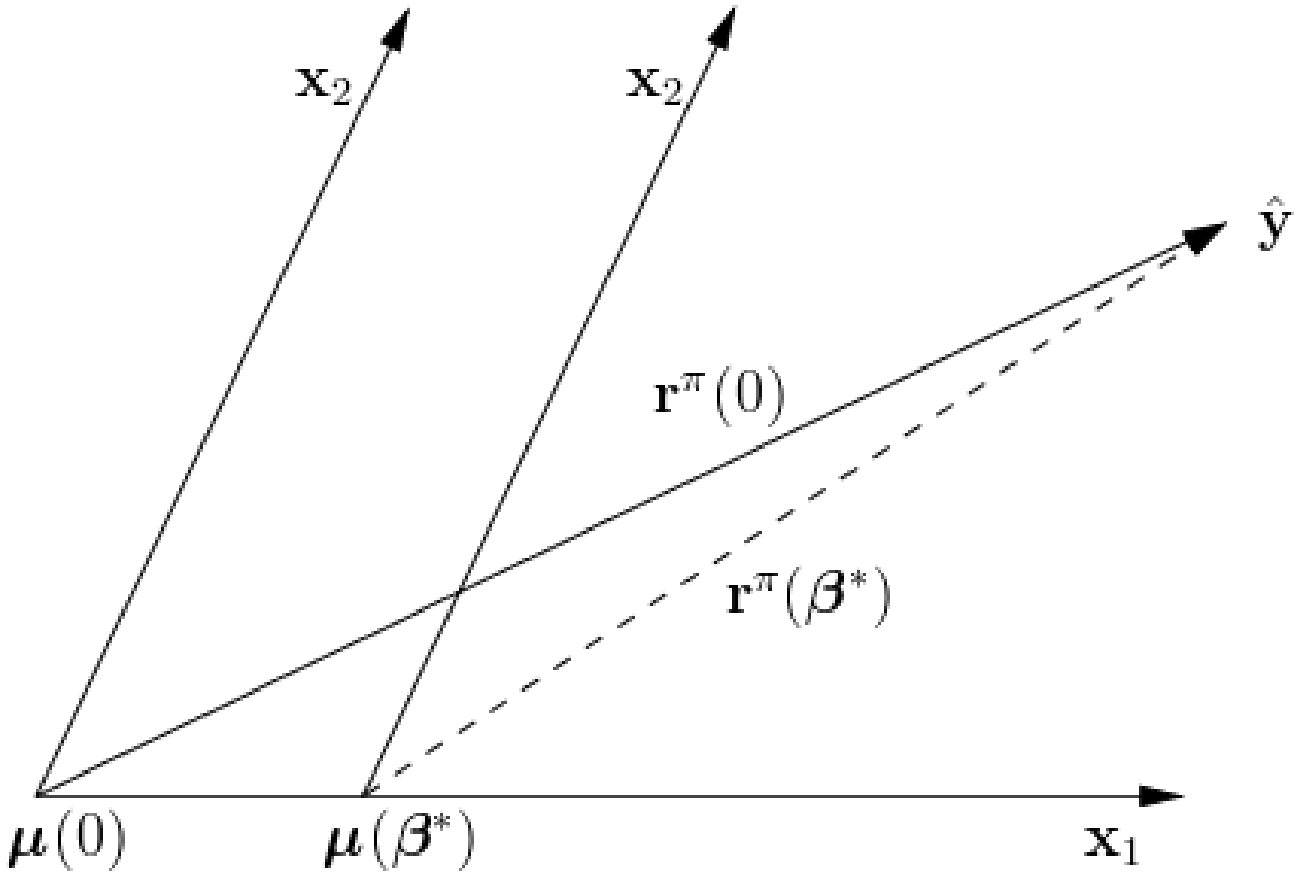
$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

By location and scale transformations we can always assume that the covariates are standardized with mean 0 and unit length, and that the response variable has mean 0,

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{for } j = 1, 2, \ldots, p.$$

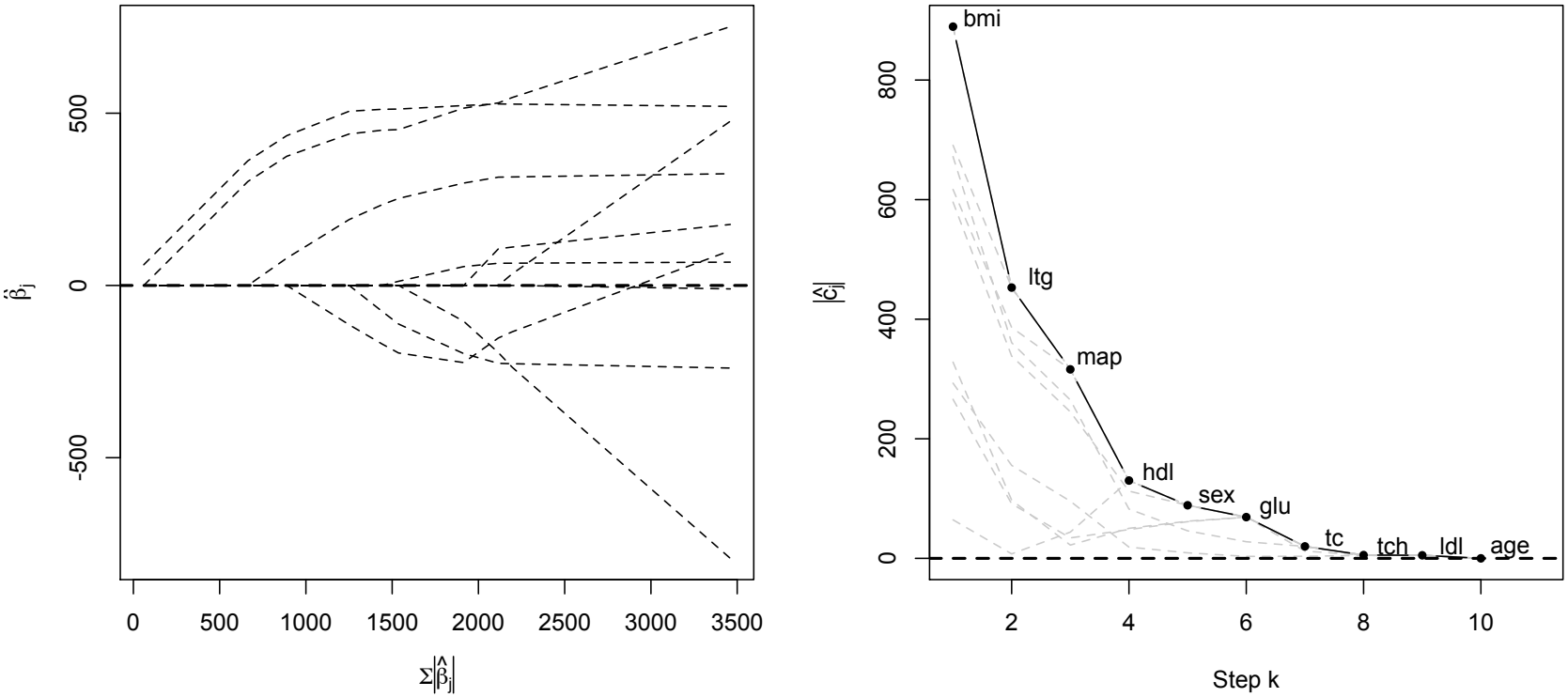
rijksuniversiteit groningen

Graphical definition of Least Angle Regression



LAR algorithm in the case of $p = 2$ covariates.
The *entire* LAR sequence require $O(p^3 + np^2)$ computations ($p < n$).

LAR analysis of the diabetes study



(left) estimates of regression coefficients $\hat{\beta}_j$ plotted versus $\sum |\hat{\beta}_j|$; (right) absolute current correlation as function of LAR step; heavy curve shows maximum current correlation.

Relationship L_1 and LAR

Let $\hat{\boldsymbol{\beta}}$ be the Lasso solution for some λ, with

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}.$$

This means that $\hat{\beta}$ solves

$$X^t X\beta - X^t y + \lambda \text{Sign}(\beta) = 0.$$

Then it is easy to show that the sign of any nonzero coordinate $\hat{\beta}_j$ must agree with the sign s_j of the current correlation $\hat{c}_j = \mathbf{x}_j^T(\mathbf{y} - \hat{\boldsymbol{\mu}})$,

$$\text{sign}(\hat{\beta}_j) = \text{sign}(\hat{c}_j) = s_j. \qquad (1)$$

Therefore, a simple modification LAR produces Lasso estimates:

> *Remove variable from active LAR set a.s.a. (1) is violated.*

IV. Generalized linear models

$\mathbf{Y}$ is a random variable with pdf

$$p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta},\lambda) = a(\mathbf{y};\lambda)\exp\{\lambda(\mathbf{y}^T\boldsymbol{\theta} - k(\boldsymbol{\theta}))\}, \quad \mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^n, \quad (2)$$

with respect to a σ-finite measure ν on $\mathbb{R}^n$, where $a(\cdot)$ and $k(\cdot)$ are specific given functions.

GLMs (McCullagh and Nelder, 1989):

$$\begin{aligned} E(Y) &= \mu \\ \eta &= \mathbf{x}^T\boldsymbol{\beta} \\ g(\mu) &= \eta \end{aligned}$$

In order to simplify our notation we denote

$$\boldsymbol{\mu}(\boldsymbol{\beta}) = \left(g^{-1}(\mathbf{x}_1^T\boldsymbol{\beta}), g^{-1}(\mathbf{x}_2^T\boldsymbol{\beta}), \ldots, g^{-1}(\mathbf{x}_n^T\boldsymbol{\beta})\right)^T.$$

rijksuniversiteit groningen

Example: logistic regression

Observations Y are binary and Bernoulli distributed,

$$\begin{aligned} p(y) &= p^y(1-p)^{1-y} \\ &= \exp\left(y\text{logit}(p) - \log(\frac{1}{1-p})\right) \end{aligned}$$

Therefore,

$$\theta = \text{logit}(p).$$

and

$$\begin{aligned} E(Y) &= p \\ \eta &= \mathbf{x}^T\beta \\ \eta &= \text{logit}(p) \end{aligned}$$

Let

$$\mu(\beta) = \left(\text{logit}(\mathbf{x}_1^T\beta), \text{logit}(\mathbf{x}_2^T\beta), \ldots, \text{logit}(\mathbf{x}_n^T\beta)\right)^T.$$

Differential manifold

Let $\mathcal{M}^n$ be the mean value parameter space.

Amari (1982):
$\mathcal{M}^n$ can be treated as a *n*-dimensional differentiable manifold.

Vos (1991):
$\boldsymbol{\mu}(\boldsymbol{\beta})$ is an embedding with domain $\mathcal{B}$

$$\widetilde{\mathcal{M}}^p = \{\boldsymbol{\mu} \in \mathcal{M}^n \,:\, \boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta}), \text{ with } \boldsymbol{\beta} \in \mathcal{B}\}. \tag{3}$$

The vector space

$$T_{\boldsymbol{\mu}(\boldsymbol{\beta})}\widetilde{\mathcal{M}}^p = \text{span}\{\partial_{\beta_1}\boldsymbol{\mu}(\boldsymbol{\beta}), \partial_{\beta_2}\boldsymbol{\mu}(\boldsymbol{\beta}), \ldots, \partial_{\beta_p}\boldsymbol{\mu}(\boldsymbol{\beta})\}$$

is called *tangent space* of $\widetilde{\mathcal{M}}^p$ at $\boldsymbol{\mu}(\boldsymbol{\beta})$.

Metric space using the Fisher information

The expected Fisher information matrix defines an inner product, denoted by $\langle , \rangle_{\mu(\beta)}$, on each tangent space:

$$\begin{aligned} \langle \mathbf{v}_1, \mathbf{v}_2 \rangle_{p(\mu)} &= E_\mu \left(\sum_{i=1}^n d\mu_{i,1} \partial_i \ell(\boldsymbol{\mu}) \cdot \sum_{j=1}^n d\mu_{2,j} \partial_j \ell(\boldsymbol{\mu}) \right) \\ &= d\boldsymbol{\mu}_1' I(\boldsymbol{\mu}) d\boldsymbol{\mu}_2. \end{aligned}$$

A given generalized linear model can be treated as a Riemannian submanifold of $(\mathcal{M}^n, \langle , \rangle_{\mu(\beta)})$. Let r be the *tangent residual vector*

$$\mathbf{r}(\boldsymbol{\mu}(\beta)) = \sum_{i=1}^n (y_i - \mu_i(\beta)) \partial_i \ell(\boldsymbol{\mu}(\beta)) \in T_{p(\boldsymbol{\mu}(\beta))} \mathcal{S}$$

It can be shown that the MLE of β is defined as

$$r(\boldsymbol{\mu}(\hat{\beta})) \perp T_{p(\boldsymbol{\mu}(\hat{\beta}))} \mathcal{M}.$$

Angles in a GLM

We will define a solution path in parameter space and need some notation:

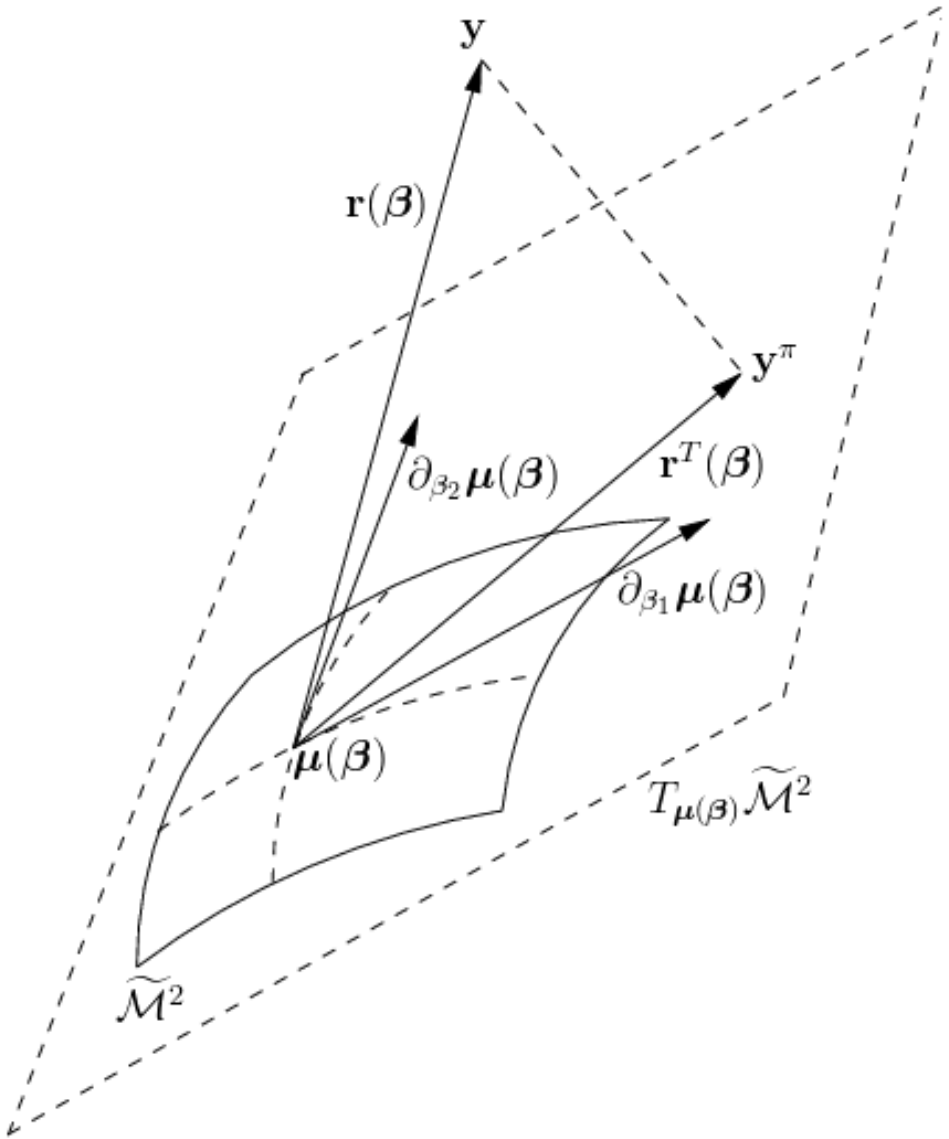
let $\boldsymbol{\beta}(\gamma)$ be a double differentiable curve;
let $\ell(\boldsymbol{\beta}(\gamma)) = \ell(\gamma)$ be the log-likelihood function;
$\partial_{\beta_m}\ell(\gamma) = \partial_m\ell(\gamma)$ be the derivative of $\ell(\gamma)$ with respect to β_m.

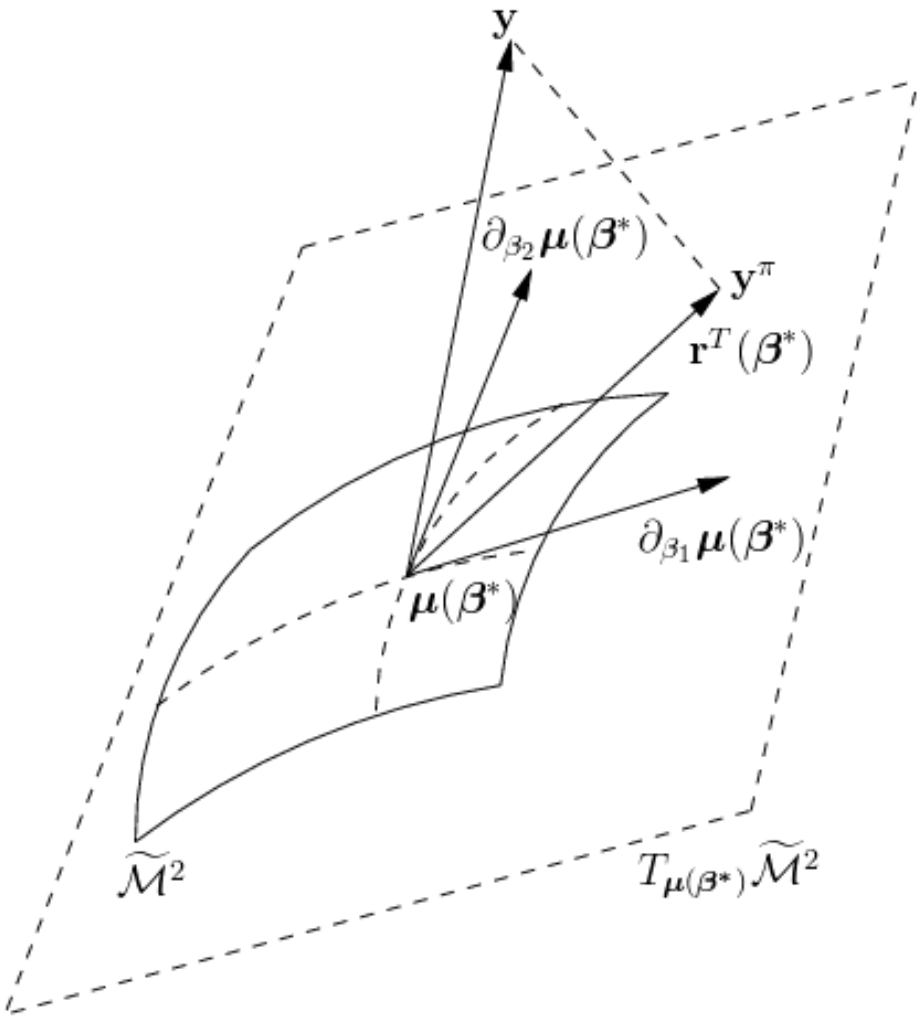
The derivative of the likelihood $\partial_m\ell(\boldsymbol{\beta}(\gamma))$ can be written as inner product between $\mathbf{r}(\boldsymbol{\beta}(\gamma))$ and m-th base of tangent space of $\mathcal{M}$ at $\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))$,

$$\partial_m\ell(\boldsymbol{\beta}(\gamma)) = \left\langle \partial_m\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma)); \mathbf{r}(\boldsymbol{\beta}(\gamma)) \right\rangle_{\boldsymbol{\mu}(\beta(\gamma))}. \tag{4}$$

Using the law of cosines, this is equivalent with

$$\begin{aligned} \partial_m\ell(\boldsymbol{\beta}(\gamma)) &= \cos\left(\rho_m(\boldsymbol{\beta}(\gamma))\right) \cdot \|\mathbf{r}\left(\boldsymbol{\beta}(\gamma)\right)\|_{\boldsymbol{\mu}(\beta(\gamma))} \cdot \|\partial_m\boldsymbol{\mu}(\beta(\gamma))\|_{\boldsymbol{\mu}(\beta(\gamma))} \\ &= \cos\left(\rho_m(\boldsymbol{\beta}(\gamma))\right) \cdot \|\mathbf{r}\left(\boldsymbol{\beta}(\gamma)\right)\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))} \cdot i_m^{1/2}(\boldsymbol{\beta}(\gamma)) \end{aligned} \tag{5}$$


$\mathbf{y}$
$\mathbf{r}(\boldsymbol{\beta})$
$\mathbf{y}^{\pi}$
$\partial_{\beta_2}\boldsymbol{\mu}(\boldsymbol{\beta})$
$\mathbf{r}^{T}(\boldsymbol{\beta})$
$\partial_{\beta_1}\boldsymbol{\mu}(\boldsymbol{\beta})$
$\boldsymbol{\mu}(\boldsymbol{\beta})$
$T_{\boldsymbol{\mu}(\boldsymbol{\beta})}\widetilde{\mathcal{M}}^2$
$\widetilde{\mathcal{M}}^2$


$\mathbf{y}$
$\partial_{\beta_2}\boldsymbol{\mu}(\boldsymbol{\beta}^*)$
$\mathbf{y}^\pi$
$\mathbf{r}^T(\boldsymbol{\beta}^*)$
$\partial_{\beta_1}\boldsymbol{\mu}(\boldsymbol{\beta}^*)$
$\boldsymbol{\mu}(\boldsymbol{\beta}^*)$
$\widetilde{\mathcal{M}}^2$
$T_{\boldsymbol{\mu}(\boldsymbol{\beta}^*)}\widetilde{\mathcal{M}}^2$

Extension of LAR (1)

Using expression (4) we have the following differential geometric identity

$$\rho_m(\boldsymbol{\beta}(\gamma)) \quad = \quad \arccos \frac{\partial_m \ell(\gamma)}{\|\mathbf{r}\left(\boldsymbol{\beta}(\gamma)\right)\|_{\boldsymbol{\mu}(\beta(\gamma))} \|\partial_m \boldsymbol{\mu}(\beta(\gamma))\|_{\boldsymbol{\mu}(\beta(\gamma))}}. \quad (6)$$

where

- $\rho_m(\boldsymbol{\beta}(\gamma))$ is the angle between $\mathbf{r}(\boldsymbol{\beta}(\gamma))$
- $\partial_m \boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))$, $i_m(\beta(\gamma))$ is the expected Fisher information for $\beta_m(\gamma)$
- $\|\cdot\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))}$ is the norm defined on the tangent space.

(6) shows that the gradient of the log-likelihood function does not generalize the equiangularity condition, since we are not considering the variation related to

$$i_m^{1/2}(\boldsymbol{\beta}(\gamma)) = \|\partial_m \boldsymbol{\mu}(\beta(\gamma))\|_{\boldsymbol{\mu}(\beta(\gamma))}.$$

rijksuniversiteit groningen

Extension of LAR (2)

Expression (6) suggests that a genuine generalization of the Least Angle Regression method should be based on the following indentity

$$r^u_m(\gamma) = i_m^{-1/2}(\boldsymbol{\beta}(\gamma)) \cdot \partial_m \ell(\gamma) = \cos\left(\rho_m(\boldsymbol{\beta}(\gamma))\right) \cdot \|\mathbf{r}(\boldsymbol{\beta}(\gamma))\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))}. \tag{7}$$

Simulation study

We have evaluated the dgLARS method by means of a simulation study.

We have simulated 1000 samples of (100,200,500) observations with 5000 variables using the following generalized linear model:

$$\begin{aligned} Y_i &\sim \text{Ber}(\pi_i), \\ \text{logit}(\pi_i) &= \eta_i, \\ \eta_i &= 1 + 2 \cdot x_{i1} + 3 \cdot x_{i2} + 4 \cdot x_{i3}. \end{aligned}$$

- We have used IID standard Gaussian predictors to obtain the design matrix $\mathbf{X}$.
- The size of the true model, i.e. the number of predictors with non-zero coefficients, is 3. .

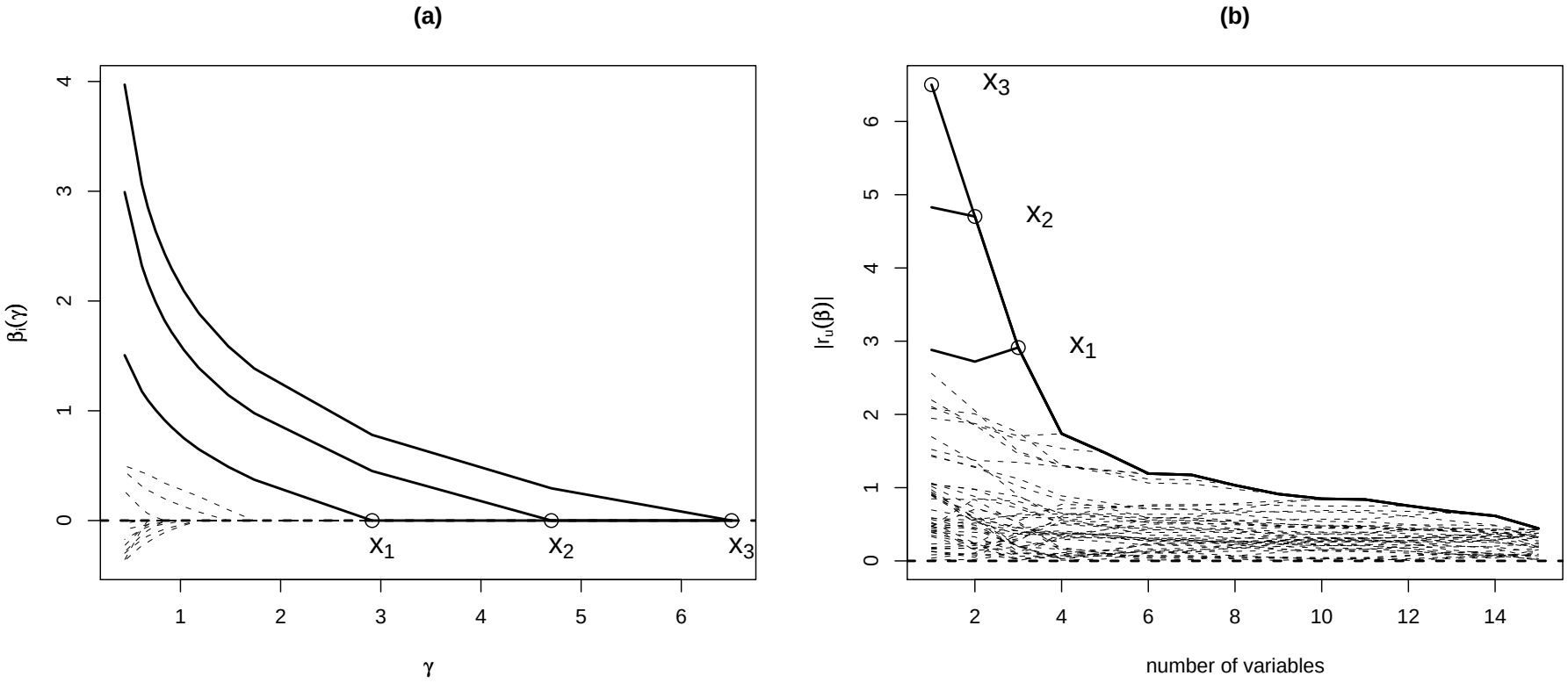


Figure (a) shows the solution path $\boldsymbol{\beta}_{\mathcal{A}}(\gamma)$ obtained with the proposed algorithm after 15 steps. Figure (b) shows the path of $|\mathbf{r}^u(\gamma)|$ as function of the number of variables included in the active set. Heavy lines identify the paths related to the variables used to simulate the response variable.

Comparison L_1 and dgLAR

200 data sets from a logistic regression model. Sample size was equal to 50/100 and the number of covariates was equal to 500/1000.
We considered a model with two important groups as follow

$$(X_1, X_2, \ldots, X_5) \backsim N_5(\mathbf{0}, \Sigma),$$
$$(X_6, X_7, \ldots, X_{10}) \backsim N_5(\mathbf{0}, \Sigma),$$
$$X_i \backsim N(0,1), \quad i = 11, \ldots, p,$$

with

$$\Sigma = \begin{pmatrix} 1 & \rho & \ldots & \rho & \rho \\ \rho & 1 & \ldots & \rho & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \ldots & 1 & \rho \\ \rho & \rho & \ldots & \rho & 1 \end{pmatrix}$$

and $\rho = 0.9$. We chose $\beta = (\underbrace{5, \ldots, 5}, \underbrace{0, \ldots, 0})$.

rijksuniversiteit groningen

	n	p	k/n 0.2	0.4	0.6	0.8	1
dgLARS	50	500	5.04	5.31	5.53	5.7	5.83
			(0.24)	(0.25)	(0.24)	(0.23)	(0.24)
	50	1000	4.92	5.23	5.44	5.62	5.80
			(0.23)	(0.24)	(0.23)	(0.23)	(0.23)
	100	500	5.95	6.29	6.47	6.66	6.80
			(0.20)	(0.20)	(0.20)	(0.20)	(0.19)
	100	1000	5.81	6.04	6.21	6.42	6.61
			(0.22)	(0.23)	(0.22)	(0.22)	(0.21)
L_1-penalty	50	500	5.04	5.26	5.21	5.2	5.15
			(0.25)	(0.24)	(0.24)	(0.23)	(0.23)
	50	1000	4.89	5.13	5.15	5.17	5.14
			(0.22)	(0.23)	(0.22)	(0.22)	(0.21)
	100	500	5.99	6.02	6.08	6.03	5.91
			(0.20)	(0.21)	(0.20)	(0.20)	(0.20)
	100	1000	5.81	5.92	5.99	5.92	5.90
			(0.22)	(0.23)	(0.22)	(0.22)	(0.22)

The numbers in parentheses are the corresponding coefficient of deviations.

Average number of true variables identified by the dgLARS and L_1-regularized logistic regression model based on 200 replicates. The average is expressed as function of the ratio between the steps of the algorithms (k) and the sample size (n).

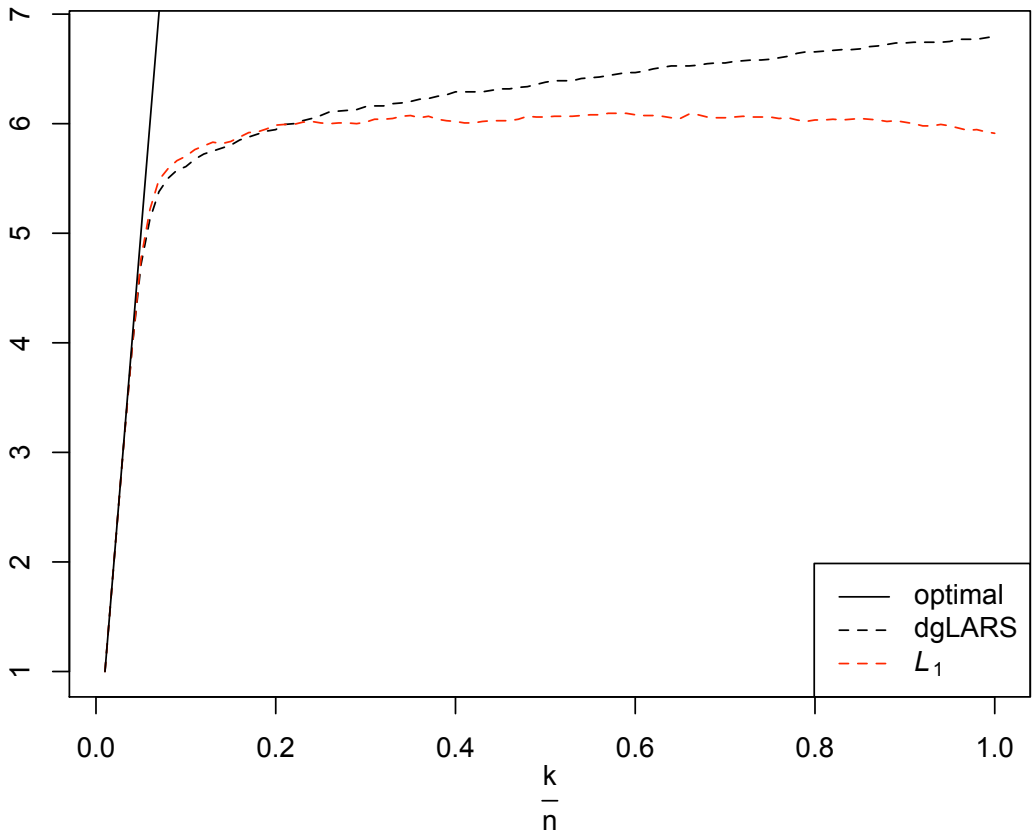


Figure shows the average number of average number of true variables identified by the dgLARS and L_1-regularized logistic regression model based on 200 replications. The average is expressed as function of the ratio between the steps of the algorithms (k) and the sample size (n).

rijksuniversiteit groningen

Conclusions:

- Penalized inference and least angle extensions are important methods to deal with high-dimensional feature spaces.
- L_1 penalized graphical models have interesting genomic applications.
- dgLAR is based on a natural generalization of the geometrical theory underlying the original LAR algorithm.

References I

[1] Allgower E. and Georg K.(1990) *Numerical Continuation Methods*, Springer, Berlin.

[2] Amari S.-I.(1985) *Differential-Geometrical Methods in Statistics (Lecture Notes in Statistics, 28)*, Springer-Verlag, New-York.

[3] Candes E. and Tao T.(2007) The Dantzig Selector: Statistical Estimation when p is much larger than n. *Ann. Statist.*, **35**(6), 2313-2351.

[4] Efron B., Hastie T., Johnstone I. and Tibshirani R.(2004) Least Angle Regression. *Ann. Statist.*, **32**(2), 407-499.

[5] Fan J. and Li R.(2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *J. Amer. Statist. Assoc.*, **96**(456), 1348-1360.

[6] Fan J. and Lv J.(2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B*, **70**(5), 849-911.

[7] Frank I.E. and Friedmann J.H.(1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**(2), 109-135.

[8] Hastie, T., Tibshirani, R. and Friedmann, J.(2001) *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, Springer.

rijksuniversiteit groningen

References II

[9] Jørgensen B.(1987) Exponential Dispersion Models (with discussion). *J. R. Statist. Soc. B*, **49**, 127-162.

[10] McCullagh P. and Nelder J.(1989) *Generalized Linear Models*, Chapman & Hall.

[11] Park M.Y. and Hastie T.(2007) L_1-regularization path algorithm for generalized linear models. *J. R. Statist. Soc. B*, **69**(4), 659-677.

[12] Rao C.R.(1945) On the distance between two populations. *Sankhya*, **9**, 246-248 .

[13] Tibshirani, R.(1996) Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc. B*, **58**(1), 267-288.

[14] Vos P.W.(1991) A Geometric Approach to Detecting Influential Cases. *Ann. Statist.*, **19**(3), 1570-1581.

[15] Witten, D.M. and Tibshrani, R (2009) Covariance-regularized regression and classification for high-dimensional problems. *J. R. Statist. Soc. B*, **71**(3), 615-636.

[16] Zou, H. and Hastie, T.(2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, **67**(2), 301-320.