Sparse survival models in high-throughput cancer studies

Ernst Wit

with Hassan Pazira, Luigi Augugliaro, Javier González, Fentaw Abegaz

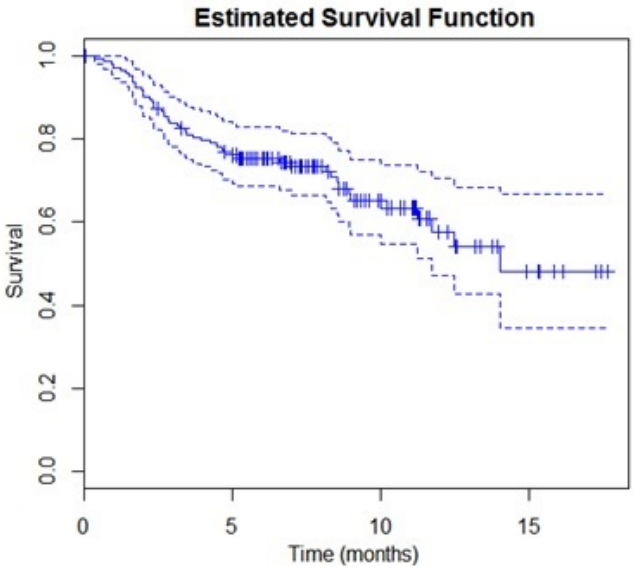
University of Groningen
Johann Bernoulli Institute of Mathematics and Computer Science
Probability and Statistics

4 November 2016

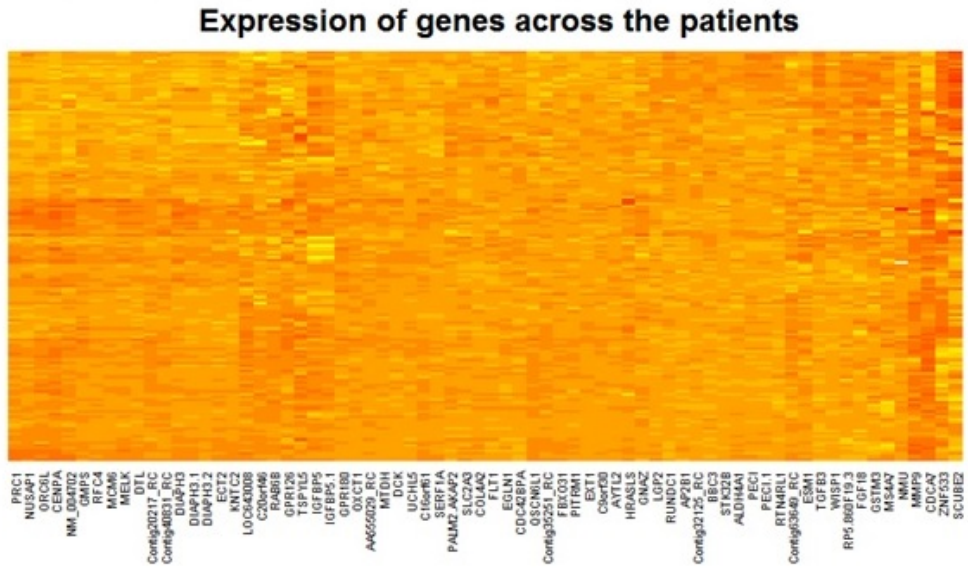
Motivation

Motivation

- 240 patients with (censored) survival times after cancer is detected.
- Explanatory variables: expression of 8000 genes in these patients.



?

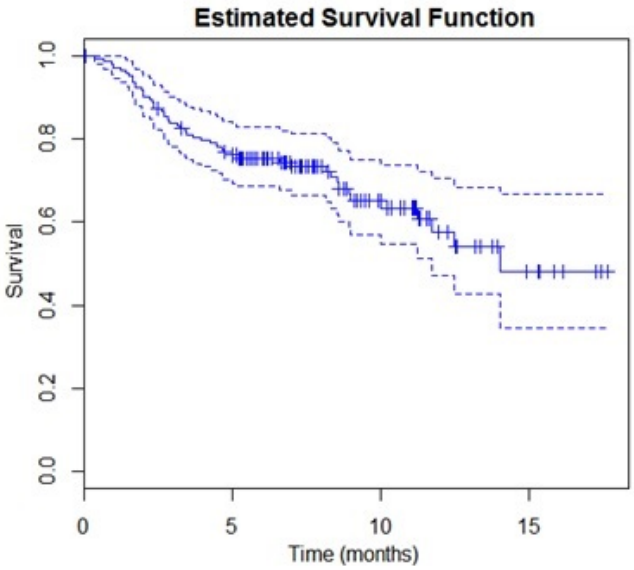


It is possible to explain the survival of the patients using the expression of the genes?

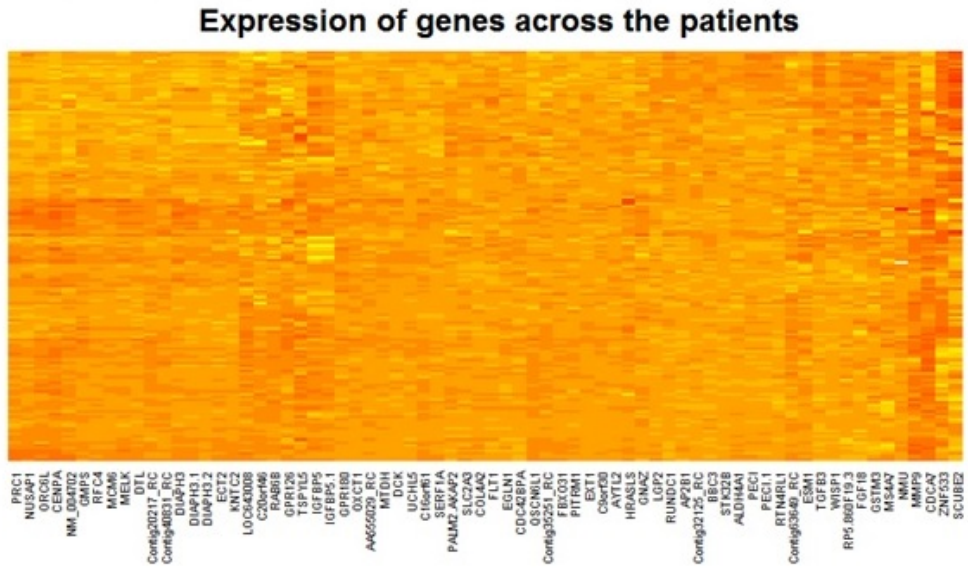
rijksuniversiteit groningen

Motivation

- 240 patients with (censored) survival times after cancer is detected.
- Explanatory variables: expression of 8000 genes in these patients.



?



It is possible to explain the survival of the patients using the expression of the genes?

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 *statistically* as a

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

Example: deletion/amplification of DNA

Problem: deletion/amplification of DNA play role in severity of breast cancer.

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

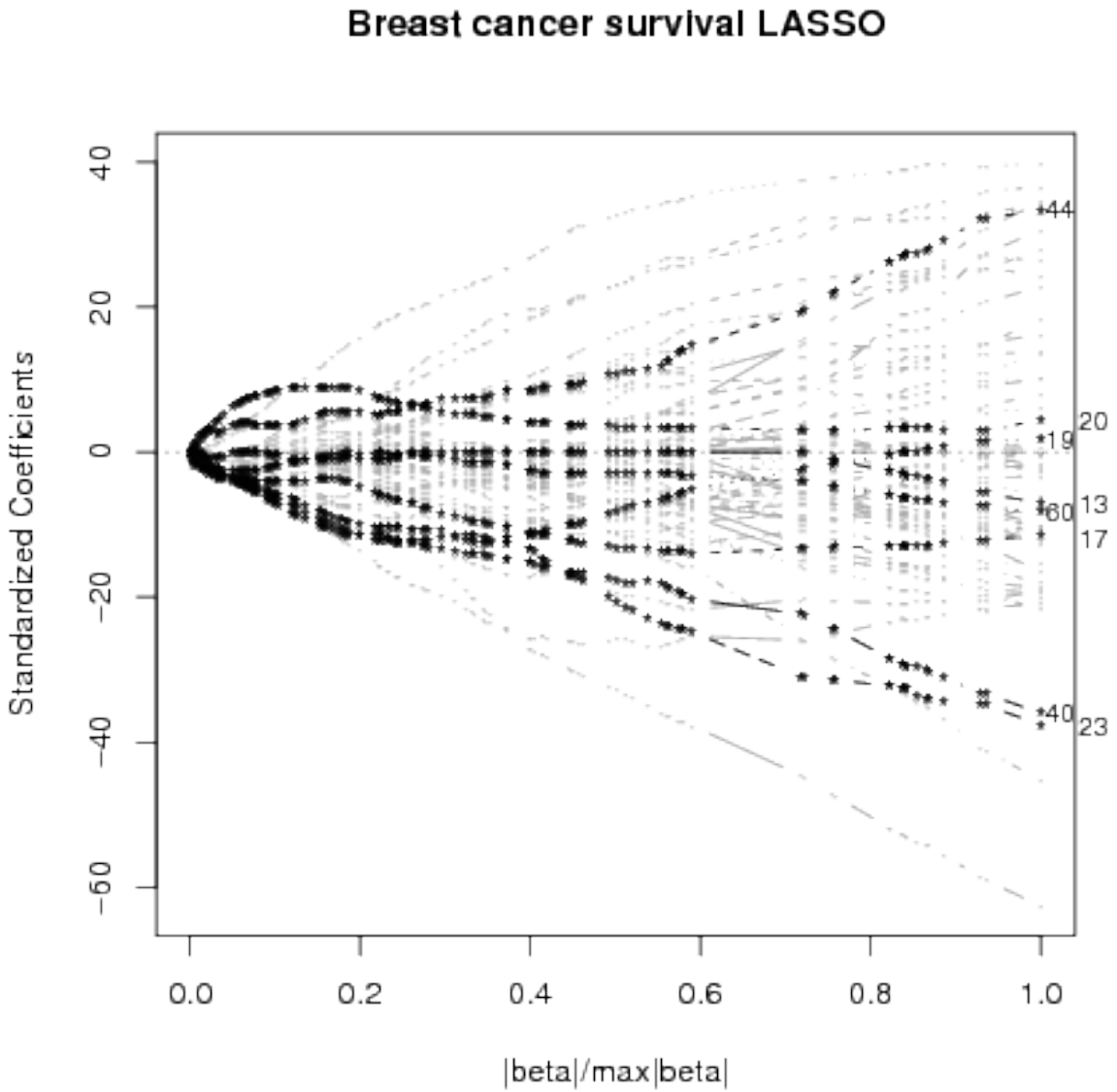
Expectation: few genes affect severity cancer (measured as NPI).

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

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

rijksuniversiteit groningen

Lasso applied to DNA deletion/amplification data

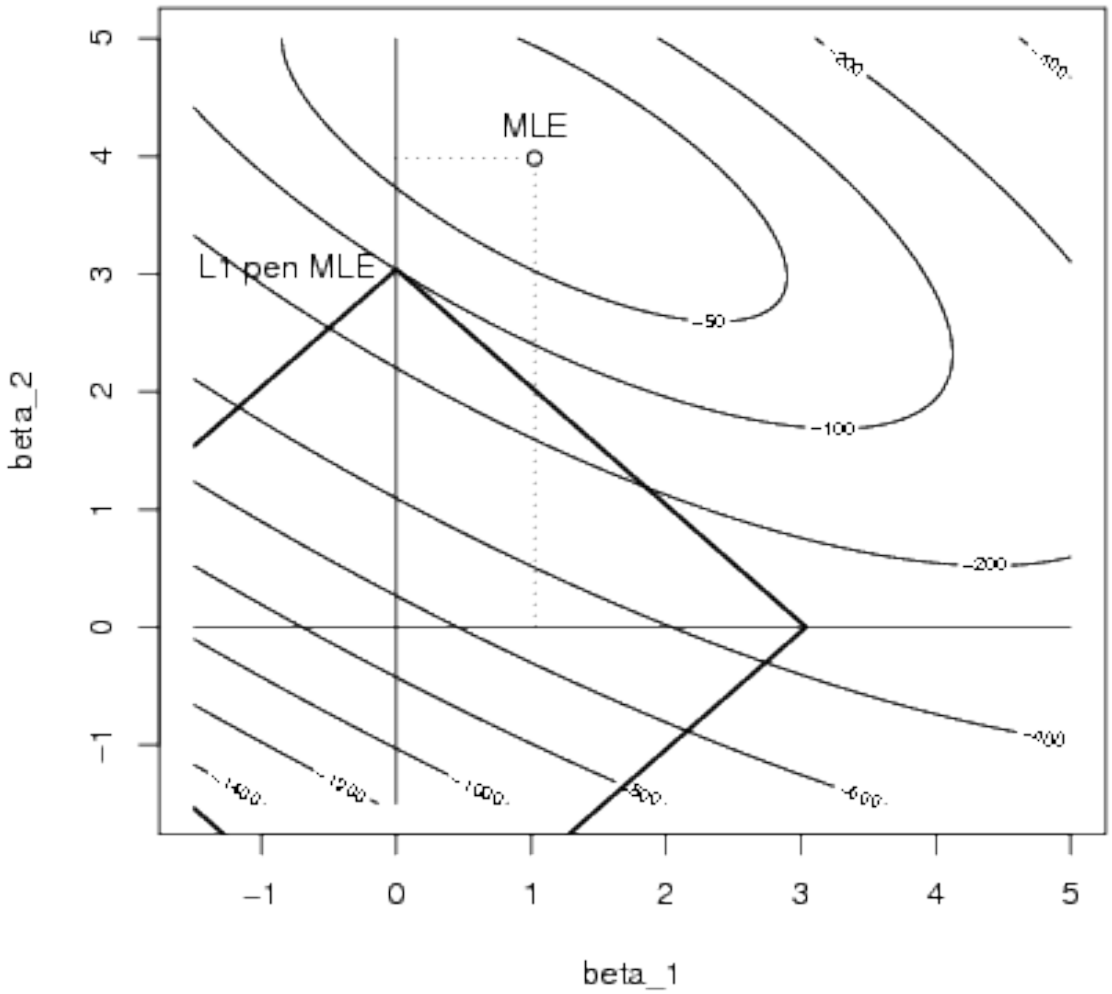


Some stopping rule selects 7 out of 59 genes.

rijksuniversiteit groningen

Geometry of the L_1 penalty = Sparsity

Maximize $\ell(\beta) = -(y - X\beta)^t(y - X\beta)$, subject to $||\beta||_1 \leq 3$.



Advantages and disadvantages of L_1 penalty

Advantages:

- *Convenient* operationalizing of sparsity.
- Convex optimization.
- Useful shrinkage behaviour with good predictive properties.

Disadvantages:

- *Arbitrary* implementation of sparsity
  based on coincidence geometry of likelihood and penalty.
- *Not invariant* to scale transformations.

Principled idea: Likelihood Sparsity

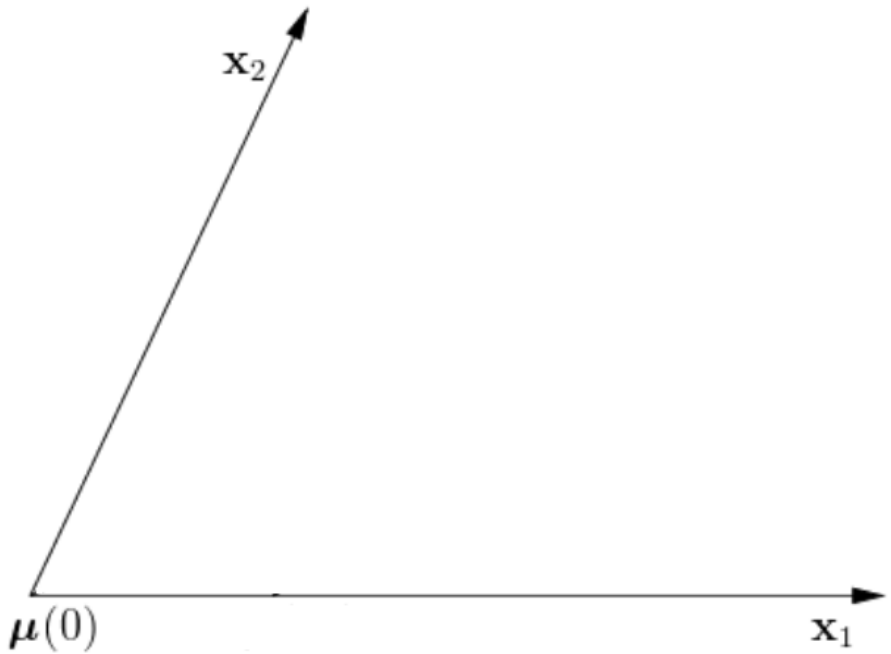
Let $\beta(\gamma_0) = \mathbf{0}$ or $\beta(\gamma_0) = (\hat{\beta}_0, \mathbf{0})$ be a sparse starting point.

Idea: define sparse solution path $\beta(\gamma)$

- as to increase ℓ as fast as possible
- independent from scale (work with angles, rather than vectors).

Needed:

Angles ρ_m for each non-active variable x_m at current $f_\beta \in \mathcal{M}$



rijksuniversiteit groningen

Principled idea: Likelihood Sparsity

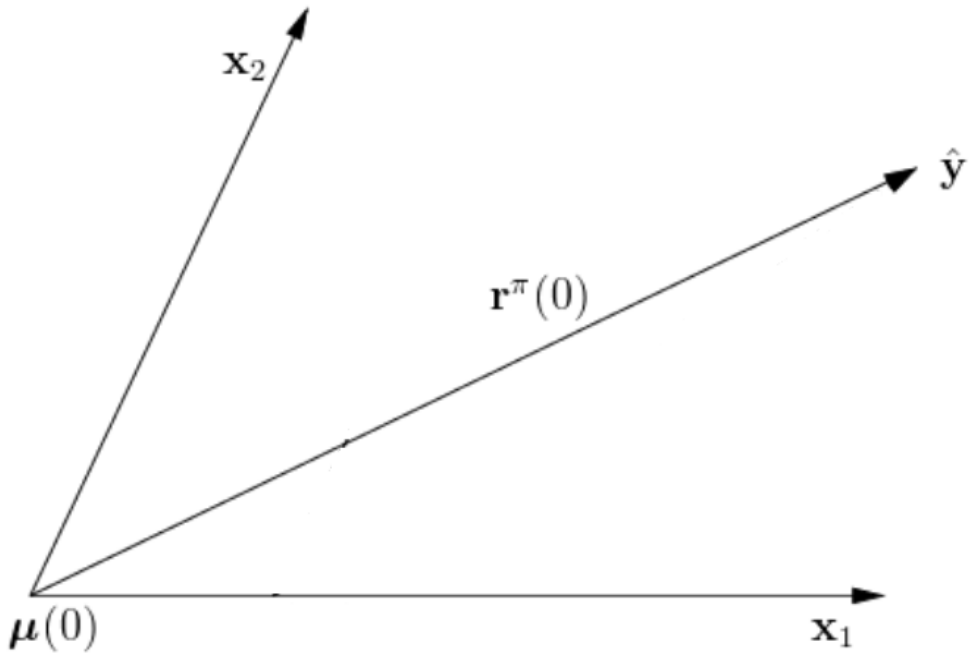
Let $\beta(\gamma_0) = \mathbf{0}$ or $\beta(\gamma_0) = (\hat{\beta}_0, \mathbf{0})$ be a sparse starting point.

Idea: define sparse solution path $\beta(\gamma)$

- as to increase ℓ as fast as possible
- independent from scale (work with angles, rather than vectors).

Needed:

Angles ρ_m for each non-active variable x_m at current $f_\beta \in \mathcal{M}$



Principled idea: Likelihood Sparsity

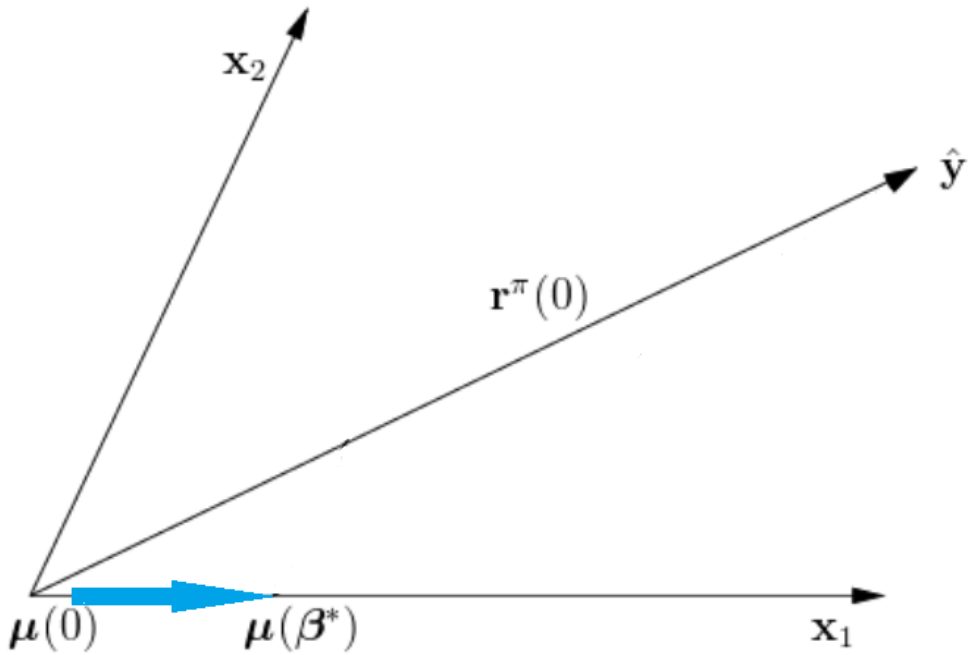
Let $\beta(\gamma_0) = \mathbf{0}$ or $\beta(\gamma_0) = (\hat{\beta}_0, \mathbf{0})$ be a sparse starting point.

Idea: define sparse solution path $\beta(\gamma)$

- as to increase ℓ as fast as possible
- independent from scale (work with angles, rather than vectors).

Needed:

Angles ρ_m for each non-active variable x_m at current $f_\beta \in \mathcal{M}$



Principled idea: Likelihood Sparsity

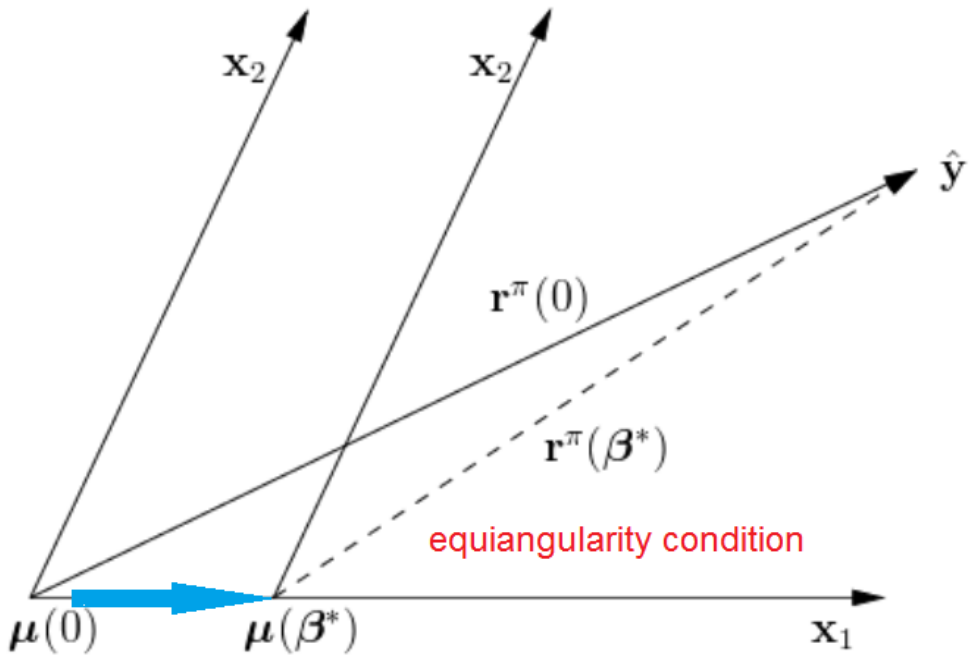
Let $\beta(\gamma_0) = \mathbf{0}$ or $\beta(\gamma_0) = (\hat{\beta}_0, \mathbf{0})$ be a sparse starting point.

Idea: define sparse solution path $\beta(\gamma)$

- as to increase ℓ as fast as possible
- independent from scale (work with angles, rather than vectors).

Needed:

Angles ρ_m for each non-active variable x_m at current $f_\beta \in \mathcal{M}$



Survival models

The Hazard function

- T: (absolutely) continuous random variable associated with the survival time.
- $f(t)$ probability density function of T.
- Hazard function:

$$\lambda(t) = \frac{f(t)}{1 - \int_0^t f(s)ds},$$

Interpretation: Instantaneous rate at which failures occur for subjects that are surviving until time t.

Relative Risk regression model

AIM: model hazard at any time t given expression profile X.

Suppose that $\lambda(t)$ depends on a p-dimensional vector of covariates

$$\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))^T.$$

Relative risk regression models

$$\lambda(t; \mathbf{x}) = \lambda_0(t)\psi(\mathbf{x}(t); \beta),$$

- $\lambda_0(t)$: base hazard function at time t.
- $\beta \in \mathcal{B} \subseteq R^p$: p-dimensional vector of unknown fixed parameters
- $\psi : R \to R$: twice continuously differentiable (relative risk function).
- $\mathcal{B}$ is such that $\psi(\mathbf{x}(t); \beta) > 0$ for each $\beta \in \mathcal{B}$.

When $\psi(\mathbf{x}(t); \beta) = \exp(\beta^T \mathbf{x}(t)) \to$ Cox regression (Cox, 1972).

rijksuniversiteit groningen

Partial Likihood

- n: observations
- t_i^f: failure times $(i = 1, \ldots, n)$.
- t_i: observation times $(i = 1, \ldots, n)$.
- $\mathcal{D}$: set of indices i for which failure time is observed, i.e.,

$$t_i = t_i^f.$$

- $\mathcal{R}(t)$: the risk set, i.e.

$$\mathcal{R}(t) = \{i \mid t_i > t\}.$$

Inference on β depends on partial likelihood function (Cox, 1972):

$$\ell_p(\beta) = \prod_{i \in \mathcal{D}} \frac{\psi(\mathbf{x}_i(t_i); \beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \beta)}.$$

PS. Maximum likelihood $\hat{\beta}$ is not available when $p > n$

rijksuniversiteit groningen

Partial Likihood

- n: observations
- t_i^f: failure times $(i = 1, \ldots, n)$.
- t_i: observation times $(i = 1, \ldots, n)$.
- $\mathcal{D}$: set of indices i for which failure time is observed, i.e.,

$$t_i = t_i^f.$$

- $\mathcal{R}(t)$: the risk set, i.e.

$$\mathcal{R}(t) = \{i \mid t_i > t\}.$$

Inference on β depends on partial likelihood function (Cox, 1972):

$$\ell_p(\beta) = \prod_{i \in \mathcal{D}} \frac{\psi(\mathbf{x}_i(t_i); \beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \beta)}.$$

PS. Maximum likelihood $\hat{\beta}$ is not available when $p > n$

rijksuniversiteit groningen

From $\ell_p(\beta)$ to a GLM

Consider $i \in \mathcal{D}$ (observed failure times).
Assume $\mathbf{Y}_i = (Y_{ih})_{h \in \mathcal{R}(t_i)}$ is a Multinomial random variable such that:

- Sample size equal to 1.
- Cell probabilities $\boldsymbol{\pi}_i = (\pi_{ih})_{h \in \mathcal{R}(t_i)} \in \Pi_i$.
- i.e. $p(\mathbf{y}; \boldsymbol{\pi}_i) = \prod_{h \in \mathcal{R}(t_i)} \pi_{ih}^{y_{ih}}$, where $\sum_{h \in \mathcal{R}(t_i)} y_{ih} = 1$.

Model space (for independent $\mathbf{Y}_i$)

$$\mathcal{S} = \left\{ \prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \pi_{ih}^{y_{ih}} : (\boldsymbol{\pi}_i)_{i \in \mathcal{D}} \in \bigotimes_{i \in \mathcal{D}} \Pi_i \right\}.$$

rijksuniversiteit groningen

From $\ell_p(\beta)$ to a GLM

Consider $i \in \mathcal{D}$ (observed failure times).
Assume $\mathbf{Y}_i = (Y_{ih})_{h \in \mathcal{R}(t_i)}$ is a Multinomial random variable such that:

- Sample size equal to 1.
- Cell probabilities $\boldsymbol{\pi}_i = (\pi_{ih})_{h \in \mathcal{R}(t_i)} \in \Pi_i$.
- i.e. $p(\mathbf{y}; \boldsymbol{\pi}_i) = \prod_{h \in \mathcal{R}(t_i)} \pi_{ih}^{y_{ih}}$, where $\sum_{h \in \mathcal{R}(t_i)} y_{ih} = 1$.

Model space (for independent $\mathbf{Y}_i$)

$$\mathcal{S} = \left\{ \prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \pi_{ih}^{y_{ih}} : (\boldsymbol{\pi}_i)_{i \in \mathcal{D}} \in \bigotimes_{i \in \mathcal{D}} \Pi_i \right\}.$$

rijksuniversiteit groningen

From $\ell_p(\beta)$ to a GLM

Fix the expected value of $\mathbf{Y}_i$

$$\begin{aligned} E_\beta(Y_{ih}) &= \pi_{ih}(\beta) \\ &:= \frac{\psi(\mathbf{x}_h(t_i);\beta)}{\sum_{j\in\mathcal{R}(t_i)}\psi(\mathbf{x}_j(t_i);\beta)}, \end{aligned}$$

Model space

$$\mathcal{S} = \left\{ \prod_{i\in\mathcal{D}} \prod_{h\in\mathcal{R}(t_i)} \pi_{ih}{}^{y_{ih}} : (\boldsymbol{\pi}_i)_{i\in\mathcal{D}} \in \bigotimes_{i\in\mathcal{D}} \Pi_i \right\}.$$

From $\ell_p(\beta)$ to a GLM

Fix the expected value of $\mathbf{Y}_i$

$$\begin{aligned} E_\beta(Y_{ih}) &= \pi_{ih}(\beta) \\ &:= \frac{\psi(\mathbf{x}_h(t_i);\beta)}{\sum_{j\in\mathcal{R}(t_i)}\psi(\mathbf{x}_j(t_i);\beta)}, \end{aligned}$$

Model space

$$\mathcal{M} = \left\{ \prod_{i\in\mathcal{D}} \prod_{h\in\mathcal{R}(t_i)} \left(\frac{\psi(\mathbf{x}_h(t_i);\beta)}{\sum_{j\in\mathcal{R}(t_i)}\psi(\mathbf{x}_j(t_i);\beta)} \right)^{y_{ih}} : \beta\in\mathcal{B} \right\}.$$

rijksuniversiteit groningen

From $\ell_p(\beta)$ to a GLM

$$\mathcal{M} = \left\{ \prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \left(\frac{\psi(\mathbf{x}_h(t_i); \beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \beta)} \right)^{y_{ih}} : \beta \in \mathcal{B} \right\}.$$

Let

$$y_{ih} = \begin{cases} 1 & \text{if } h = i \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \left(\frac{\psi(\mathbf{x}_h(t_i); \beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \beta)} \right)^{y_{ih}} = \prod_{i \in \mathcal{D}} \frac{\psi(\mathbf{x}_h(t_i); \beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \beta)} = \ell_p(\beta)$$

Likelihood associated $\mathcal{M}$ is equivalent to partial likelihood $\ell_p(\beta)$.

From $\ell_p(\beta)$ to a GLM

$$\mathcal{M} = \left\{ \prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \left(\frac{\psi(\mathbf{x}_h(t_i); \beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \beta)} \right)^{y_{ih}} : \beta \in \mathcal{B} \right\}.$$

Let

$$y_{ih} = \begin{cases} 1 & \text{if } h = i \\ 0 & \text{otherwise} \end{cases}$$

Then

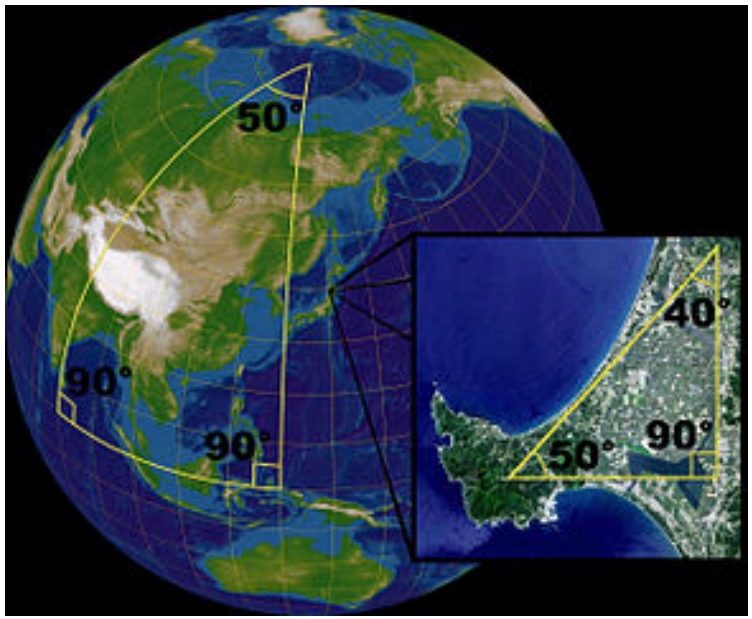
$$\prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \left(\frac{\psi(\mathbf{x}_h(t_i); \beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \beta)} \right)^{y_{ih}} = \prod_{i \in \mathcal{D}} \frac{\psi(\mathbf{x}_h(t_i); \beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \beta)} = \ell_p(\beta)$$

Likelihood associated $\mathcal{M}$ is equivalent to partial likelihood $\ell_p(\beta)$.

Bit of differential geometry

Partial likelihood $\ell_p(\beta)$ is a manifold

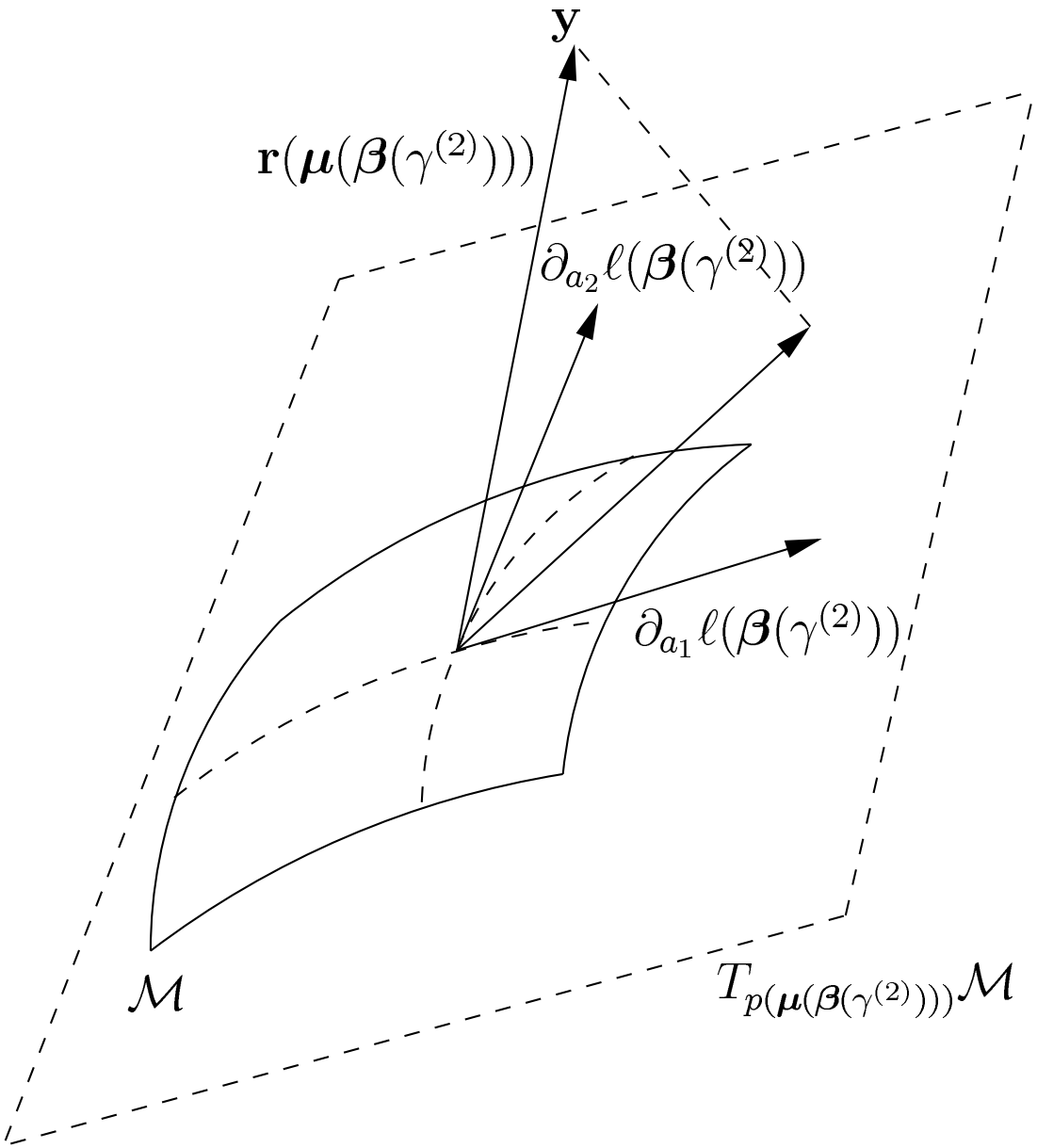
$$\mathcal{M} = \left\{ \prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \left(\frac{\psi(\mathbf{x}_h(t_i); \beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \beta)} \right)^{y_{ih}} : \beta \in \mathcal{B} \right\}.$$



- $\mathcal{M}$ is **model space** (model earth):
  Resembles locally a Euclidean space (Amari, 1982; Vos, 1991).
- Inner product is the Fisher Information.

rijksuniversiteit groningen

First view of $T_f\mathcal{M}$...



rijksuniversiteit groningen

Angles in a Relative Risk regression model

The angle ρ_m between tangent residual vector and $\partial_m \ell(\boldsymbol{\mu})$:

$$\begin{aligned} \partial_m \ell(\boldsymbol{\mu}) &= \langle \partial_m \ell(\boldsymbol{\mu}), r(\boldsymbol{\mu}) \rangle_{f_\mu} \\ &= \cos(\rho_m) \cdot \|\mathbf{r}(\boldsymbol{\mu})\|_{f_\mu} \cdot \|\partial_m \ell(\boldsymbol{\mu})\|_{f_\mu} \\ &= \cos(\rho_m) \cdot \|\mathbf{r}(\boldsymbol{\mu})\|_{f_\mu} \cdot i_m^{1/2}(\boldsymbol{\mu}), \end{aligned}$$

then we obtain:

$$\rho_m = \arccos \frac{\partial_m \ell(\boldsymbol{\mu})}{\|\mathbf{r}(\boldsymbol{\mu})\|_{f_\mu} \cdot i_m^{1/2}(\boldsymbol{\mu})}.$$

where

- ρ_m is angle between $\mathbf{r}(\boldsymbol{\mu})$ and $\partial_m \ell(\boldsymbol{\mu})$.
- $i_m(\boldsymbol{\mu})$ is the expected Fisher information for β_m
- $\|\cdot\|_{f_\mu}$ is the norm defined on $T_{f_\mu}\mathcal{F}$.

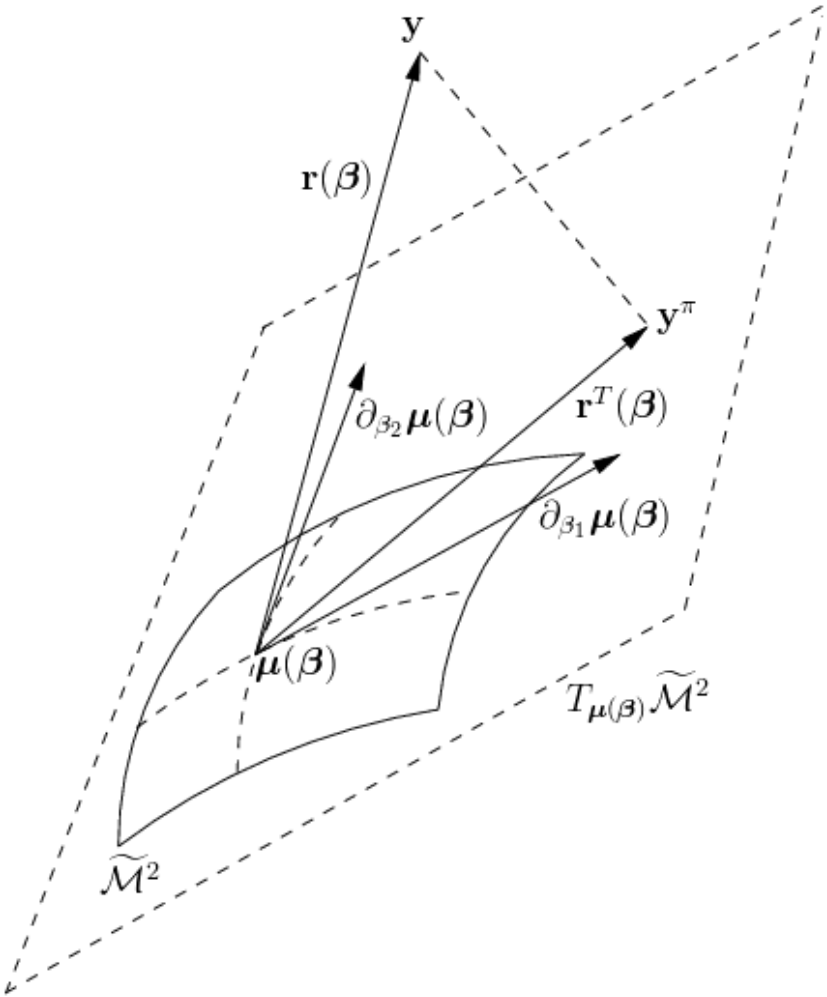
Note: Gradient of log-likelihood function is not sufficient!

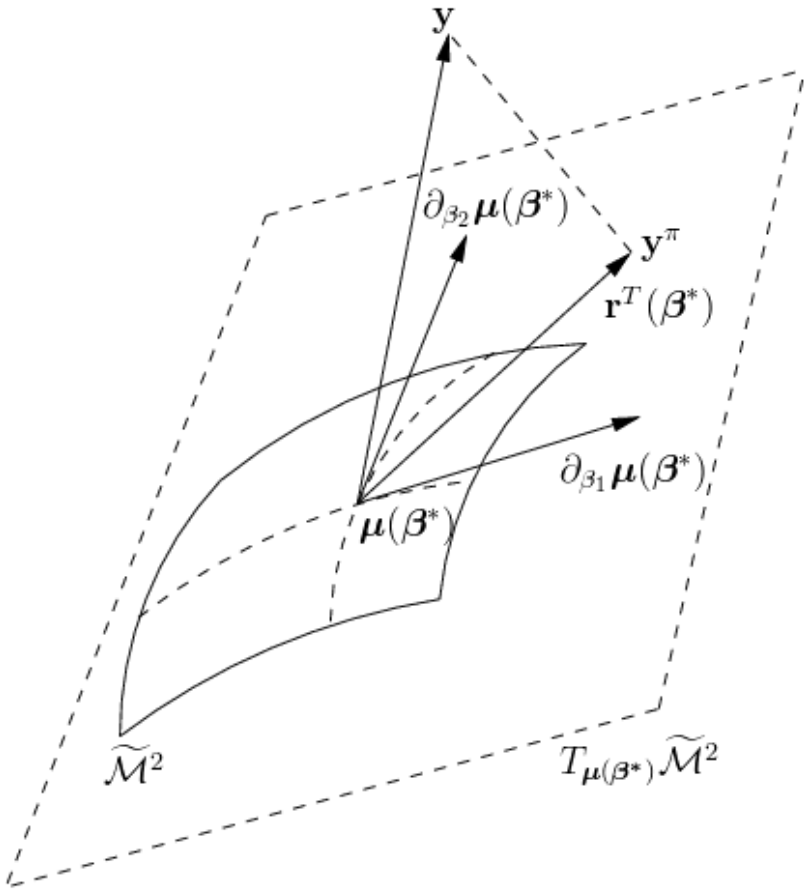
rijksuniversiteit groningen

Method (Differential geometric Relative Risk)

Step	*Algorithm*
1	start with intercept only model
2	repeat
3	increase parameters of active variables keeping angles between their scores and residual tangent vector same
4	if angle of not-included variable is same as those currently in model, include that variable in active set
5	until a stopping rule is met

Including first x_1, then x_2...





Rao score statistic

Let's return to following differential geometric identity

$$\rho_m = \arccos \frac{\partial_m \ell(\boldsymbol{\mu})}{\|\mathbf{r}(\boldsymbol{\mu})\|_{f_\mu} \cdot i_m^{1/2}(\boldsymbol{\mu})}$$

Note: $\|\mathbf{r}(\boldsymbol{\mu})\|_{f_\mu}$ does not depend on m!

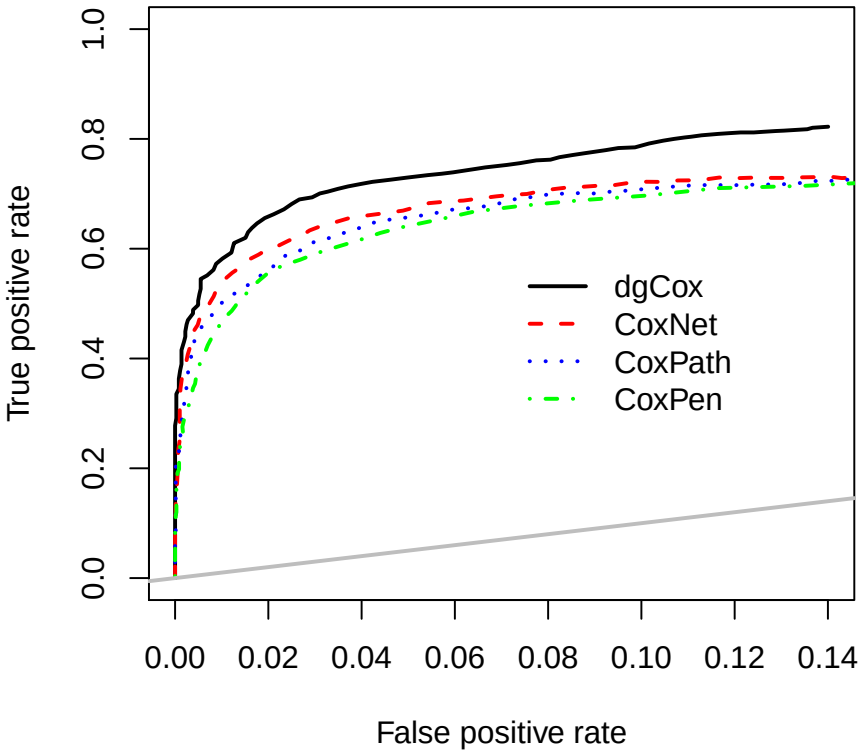
So, **equivalently** we can use **Rao score statistic**:

$$r_m^u(\boldsymbol{\beta}) = \frac{\partial_m \ell(\boldsymbol{\mu})}{i_m^{1/2}(\boldsymbol{\mu})}.$$

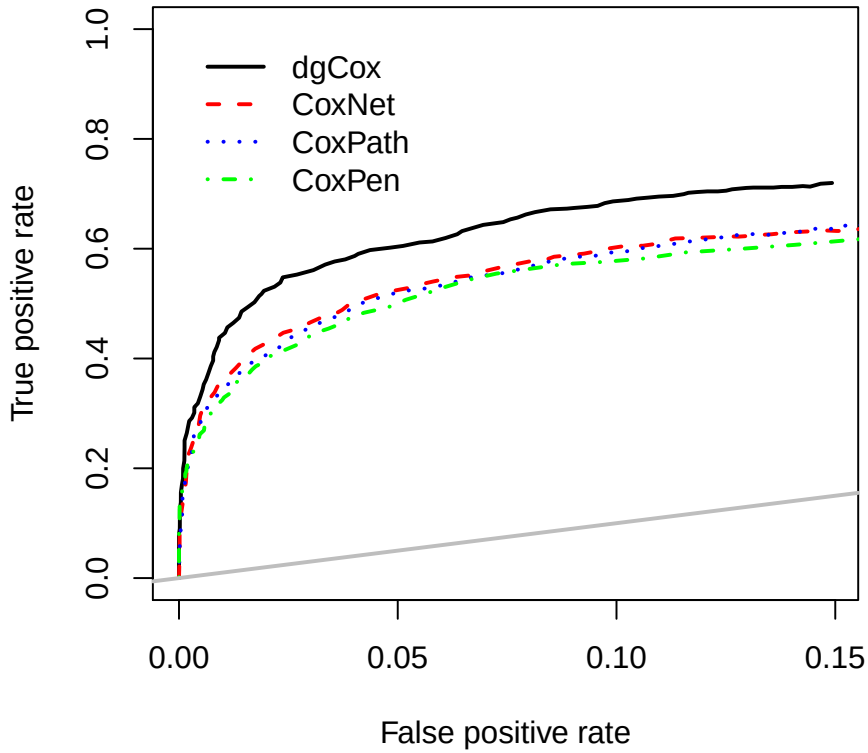
The larger $r_m^u(\boldsymbol{\beta})$, the smaller ρ_m, the better!

1. Simulation study: ROC curve

(e) $\rho = 0.9$ $\pi = 0.2$



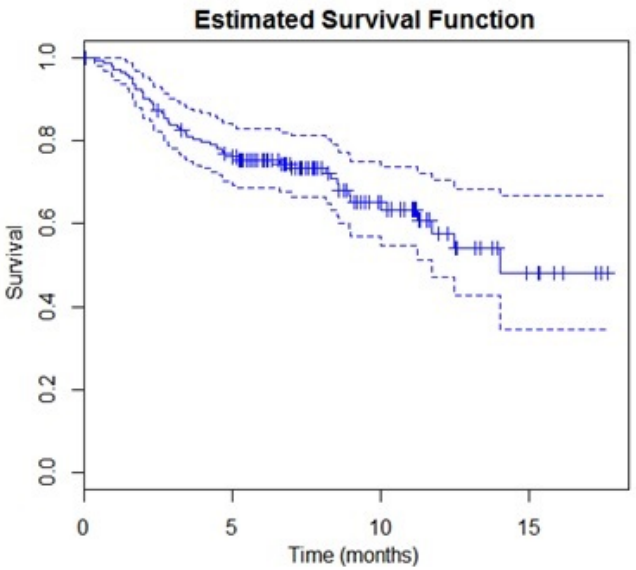
(f) $\rho = 0.9$ $\pi = 0.4$



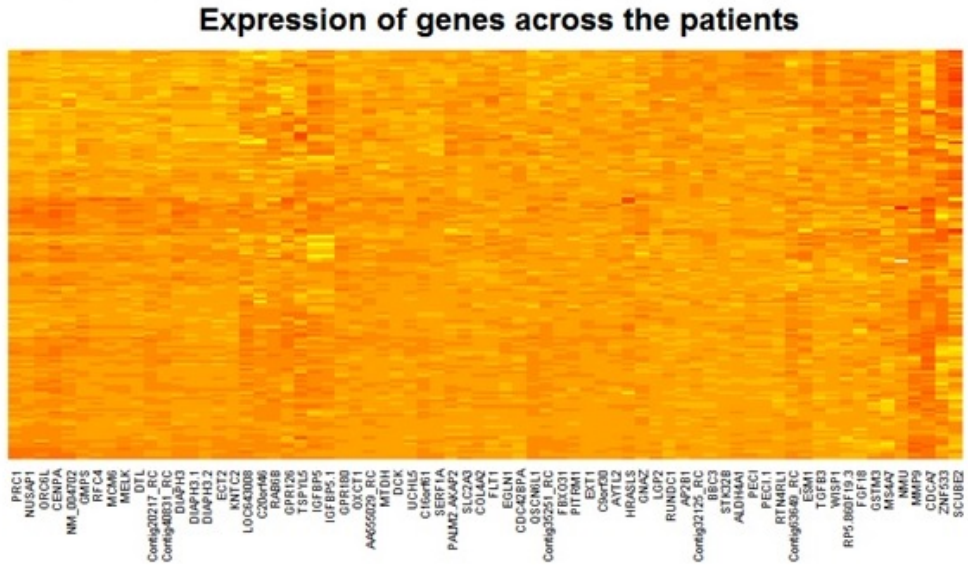
DgCox is best when there are (large) correlation among the predictors $\rightarrow$ Most real cases!

2. Diffuse large-B-cell lymphoma dataset (DLBCL)

- Survival times of 240 patients.
- Gene expression measurements on 7399 genes after chemotherapy.
- Missing data imputed via k-nearest neighbours.
- **Aim:** molecular predictor model of survival after chemo.



?



Model selection

- BIC

$$BIC(\hat{\beta}, \gamma) = -2\,\ell_p(\hat{\beta}(\gamma)) + \log(n) df\,,$$

- AIC

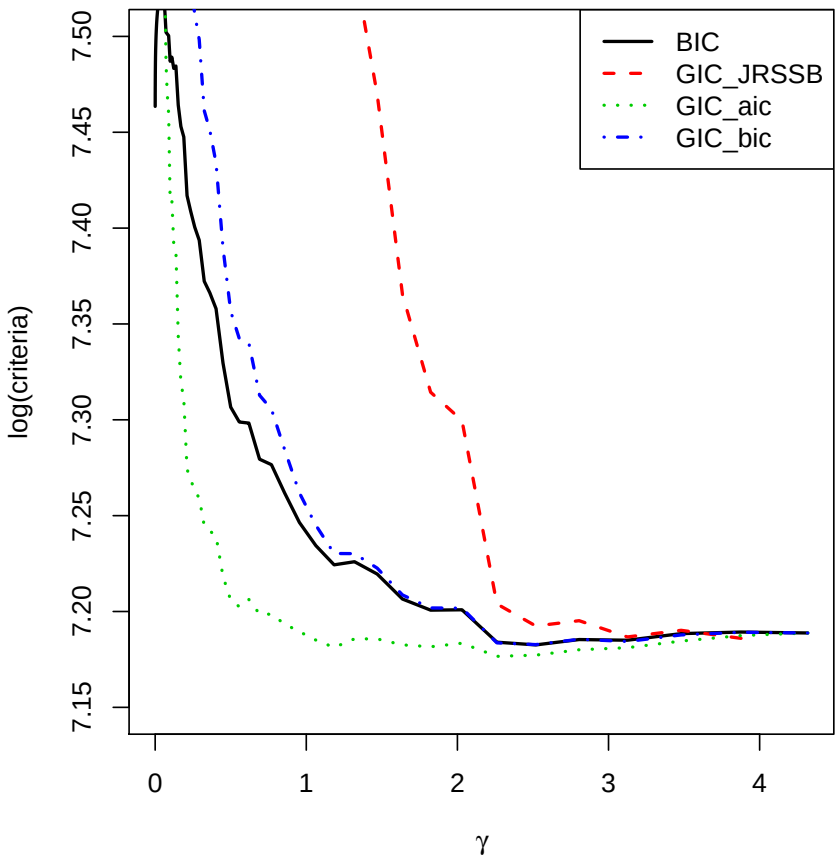
$$BIC(\hat{\beta}, \gamma) = -2\,\ell_p(\hat{\beta}(\gamma)) + 2df\,,$$

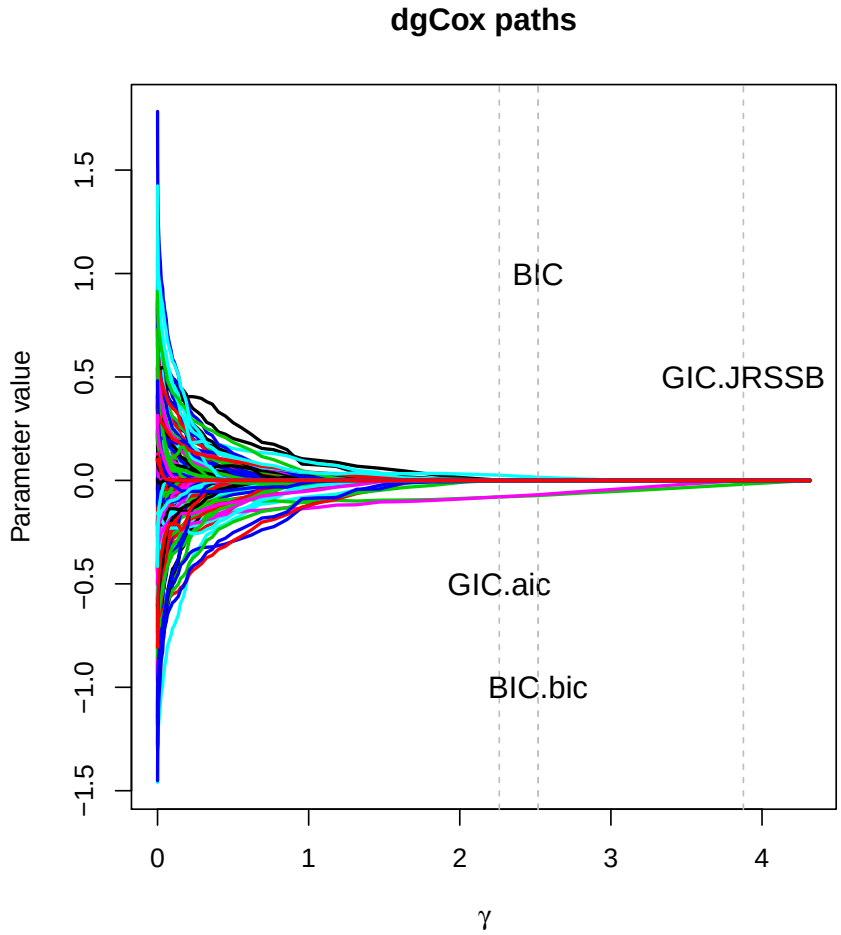
- Derivation of the GIC using M-estimators

$$GIC.aic(\hat{\beta}, \gamma) = -2\,\ell_p(\hat{\beta}(\gamma)) + \mathrm{tr}(R^{-1}(\hat{\beta}, \gamma)\, Q(\hat{\beta}, \gamma)),$$

- (Completely different) GIC proposed in Fan (2013)

Paths and selected model





Selected genes (GIC.aic)

X4131(-0.080), X5054(-0.079), X5172 (0.0012), X6321(0.0253).

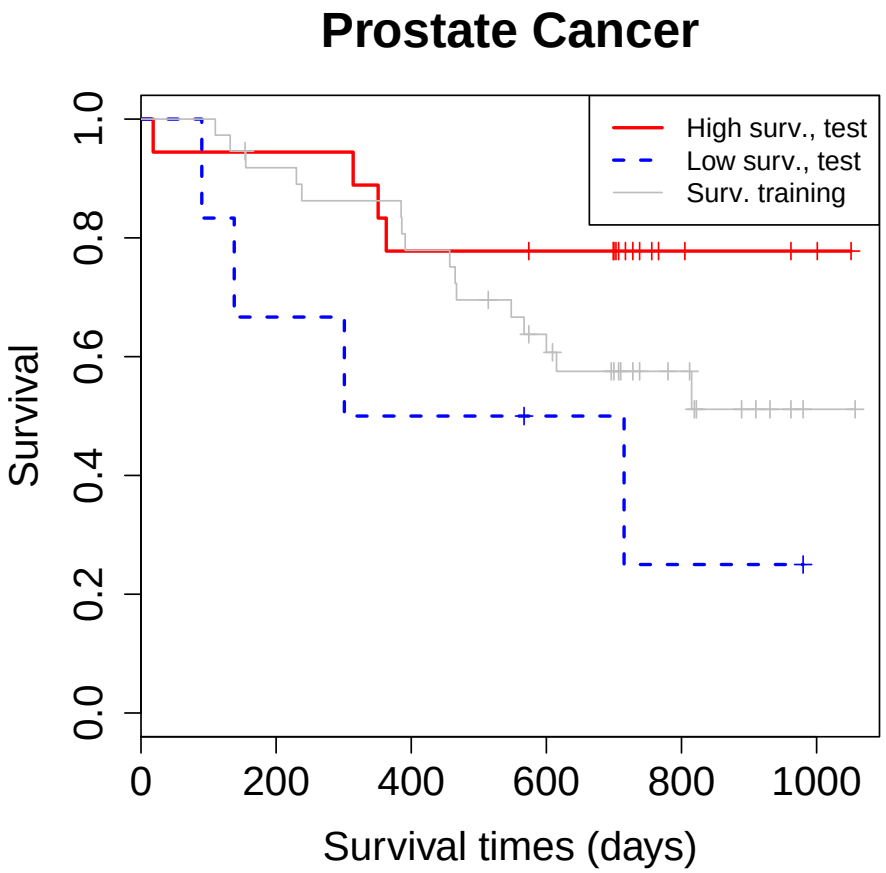
rijksuniversiteit groningen

3. Real data validation

We consider four recent studies on cancer survival:

Cancer	n	# uncen	p	# genes sel.	Ref.
Prostate	61	24	162	33	Ross et al (2012)
Ovarian	103	57	306	48	Gillet et al. (2012)
Skin	54	47	30807	44	Jonsson et al. (2010)
Colon	125	70	23698	62	Loboda et al. (2011)

Prostate cancer and Ovarian cancer survival



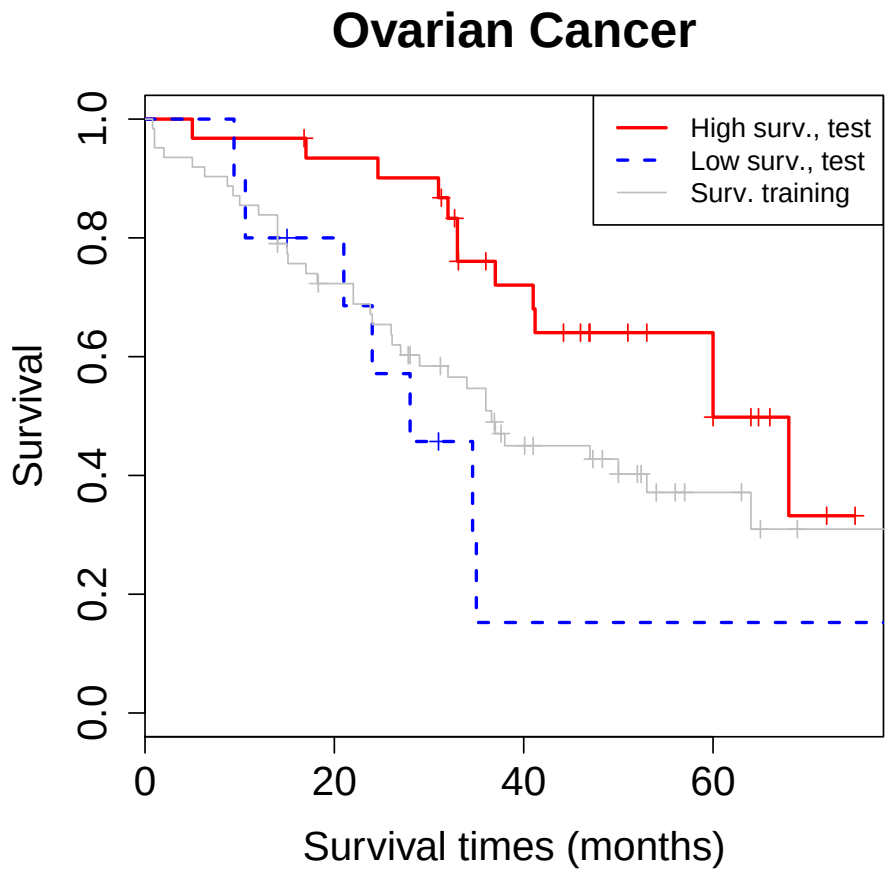
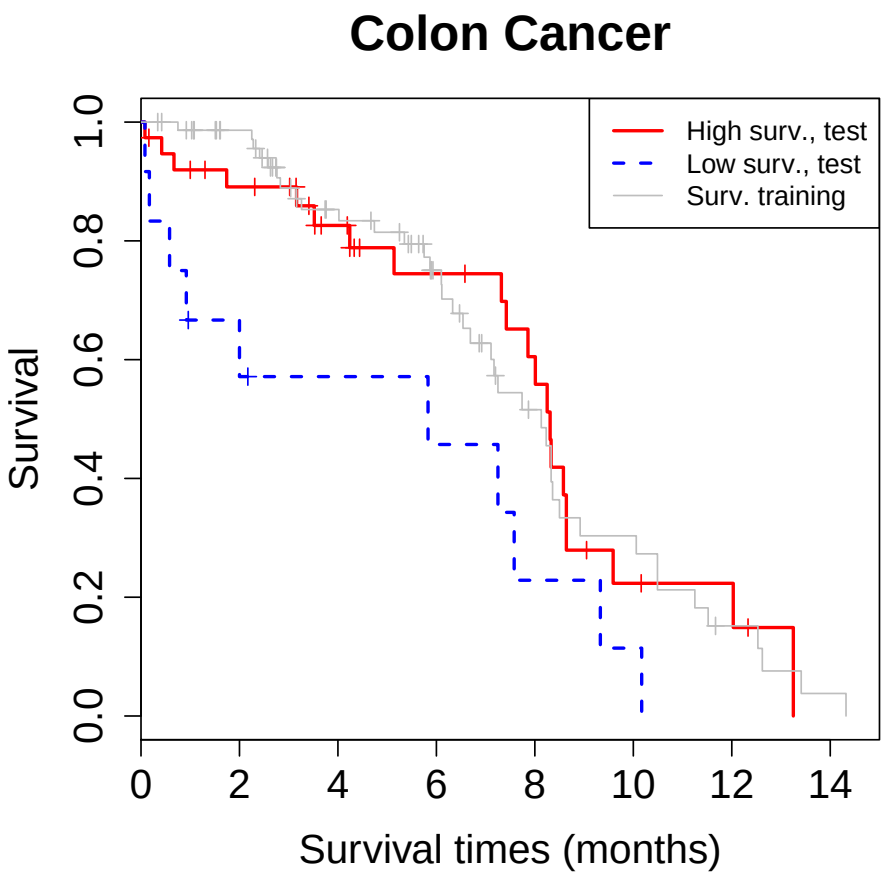


Figure: Performance on independent test data:

Prostate data: n = 61, p = 162, p_{sel} = 33, p-value = 0.03

Ovarian data: n = 103, p = 306, p_{sel} = 48, p-value = 0.01

Skin cancer and Colon cancer survival



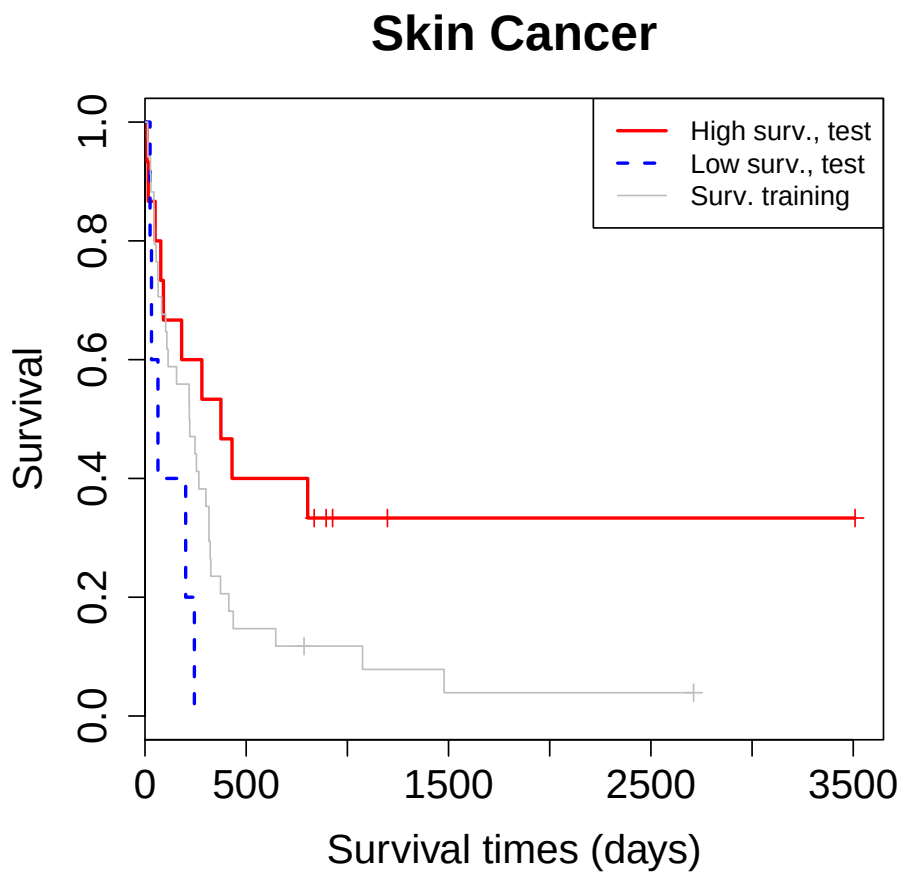


Figure: Performance on independent test data:

Skin data: n = 54, p = 30,807, p_{sel} = 44, p-value = 0.07

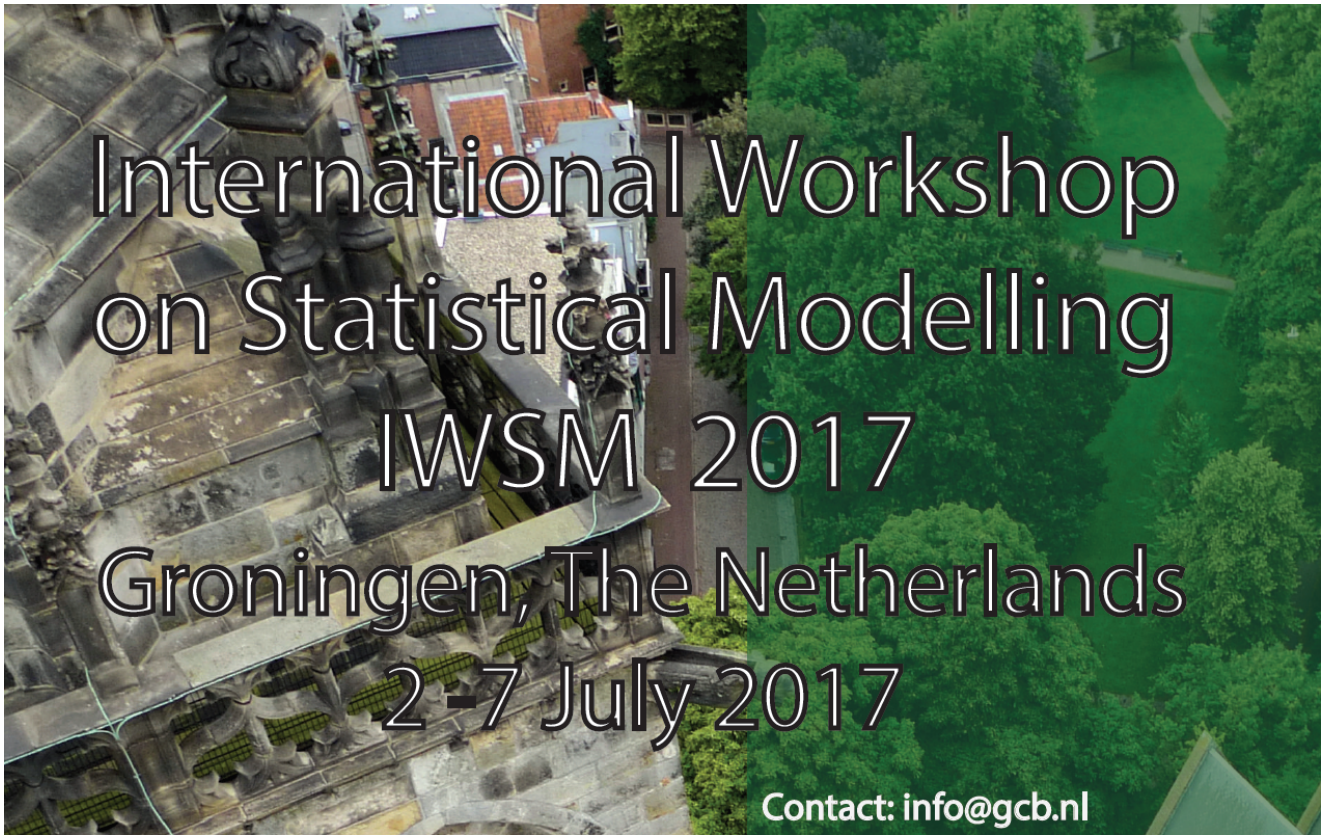
Colon data: n=125, p = 23,698, p_{sel} = 62, p-value = 0.02

Concluding remarks

- Sparse survival modelling is often *ad hoc*.
- Sparse parameter walks on survival likelihood are
  - intuitively appealing;
  - computationally feasible (Augugliaro et al. 2014, JSS);
  - theoretically sound (Augugliaro et al. 2013, JRSS-B);
  - methodologically extendible (Augugliaro et al. 2016, Biometrika);
- Our method outperforms existing sparse Cox regression techniques.

Method will soon be available in `R` from package `dglars`.

IWSM 2017: 2-7 July 2017, Groningen, the Netherlands



- Keynote speakers: Laura Sangalli, Tom Snijders, Jelle Goeman...
- No parallel sessions!
- Website: http://iwsm2017.webhosting.rug.nl/
- 3-page abstract submission before January 30, 2017

rijksuniversiteit groningen

Computational considerations

dgLARS as a Z-estimator

Assume: first k predictors and intercept included in $\mathcal{A}$.

Sparse estimator $\hat{\beta}$ is solution of following system:

$$\begin{cases} r_0^u(\beta) & = & 0 \\ r_1^u(\beta) - \gamma v_1 & = & 0 \\ \vdots & & \vdots \\ r_k^u(\beta) - \gamma v_k & = & 0 \end{cases}$$

where $v_m = \text{sign}(r_m^u(\beta))$.

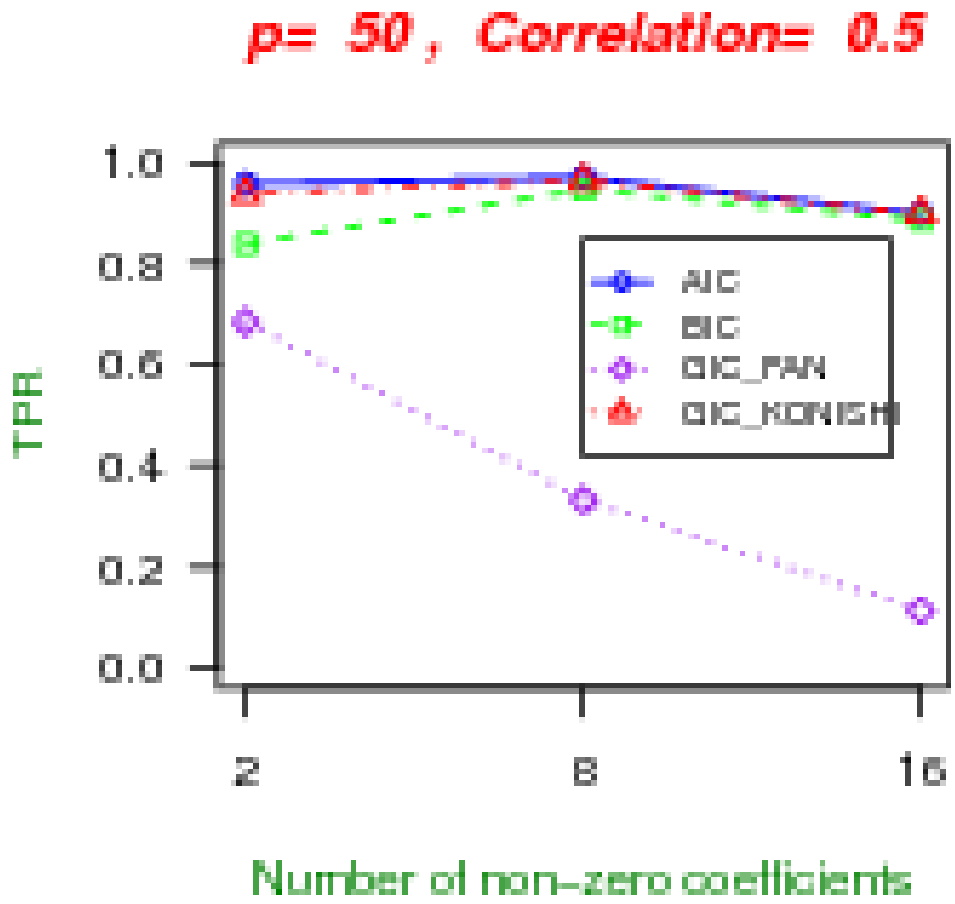
Method computes finite sequence of transition points,

$$0 \leq \gamma^{(K)} \leq \ldots \leq \gamma^{(2)} \leq \gamma^{(1)},$$

s.t. for each $\gamma^{(k)}$ one of two things can occur:

- **Inclusion condition:** $\left| r_q^u(\hat{\beta}) \right| = \gamma^{(k)}$, with $q \notin \mathcal{A}$
- **Exclusion condition:** $\text{sign}(r_m^u(\hat{\beta}))) \neq \text{sign}(\hat{\beta}_m)$ with $m \in \mathcal{A}$

Model selection



AIC/GIC have slightly liberal, but overall good performance.

rijksuniversiteit groningen