Modeling Survival and Risk

Regression techniques in survival analysis

Prof. Dr. José Manuel Sánchez-Santos

Dpto. Estadística - Universidad de Salamanca

Grupo de Bioinformática y Genómica Funcional (Lab 19)

Centro de Investigación del Cáncer (CIC-Usal)

Outline

Introduction

Regularized regression models

- Ridge
- Lasso
- Elastic-net

Survival and Risk

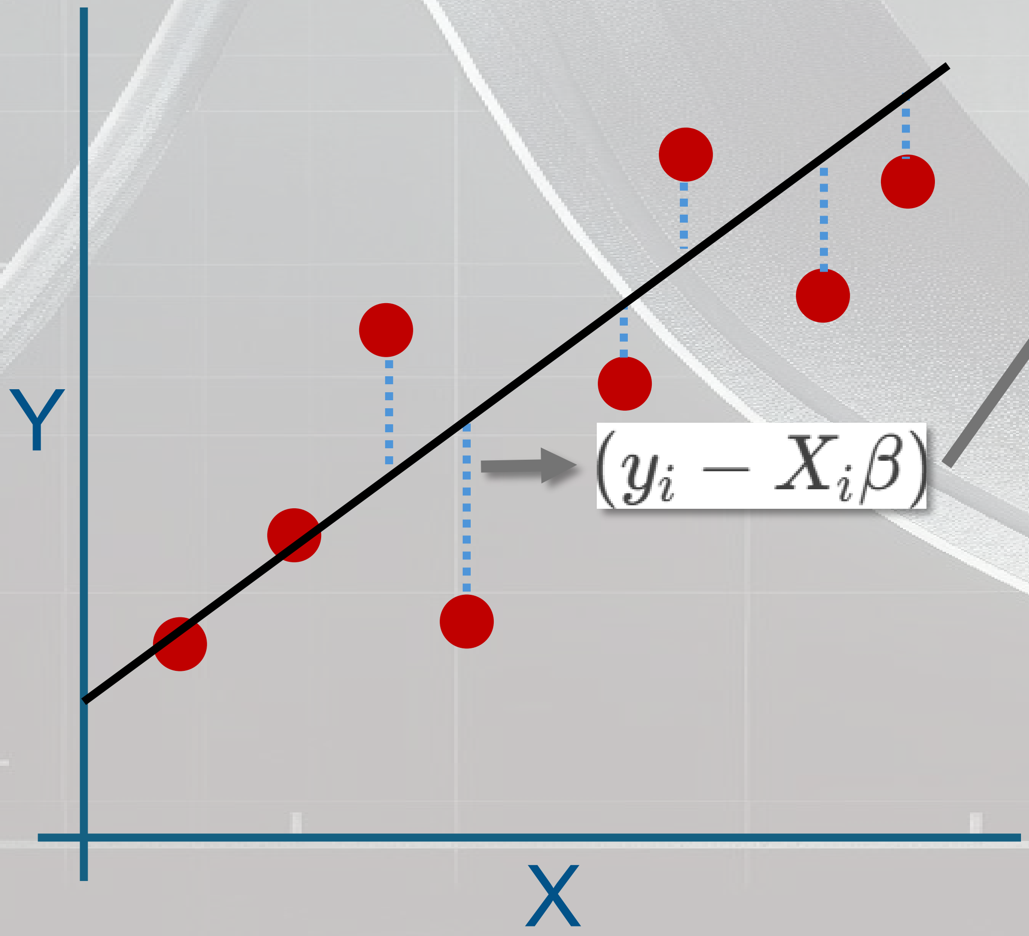
Applications

- Gene-Phenotype
- Gene-Survival
- Patient-Risk

Introduction

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

Linear regression: Ordinary Least Squares



$(y_i - X_i\beta)$

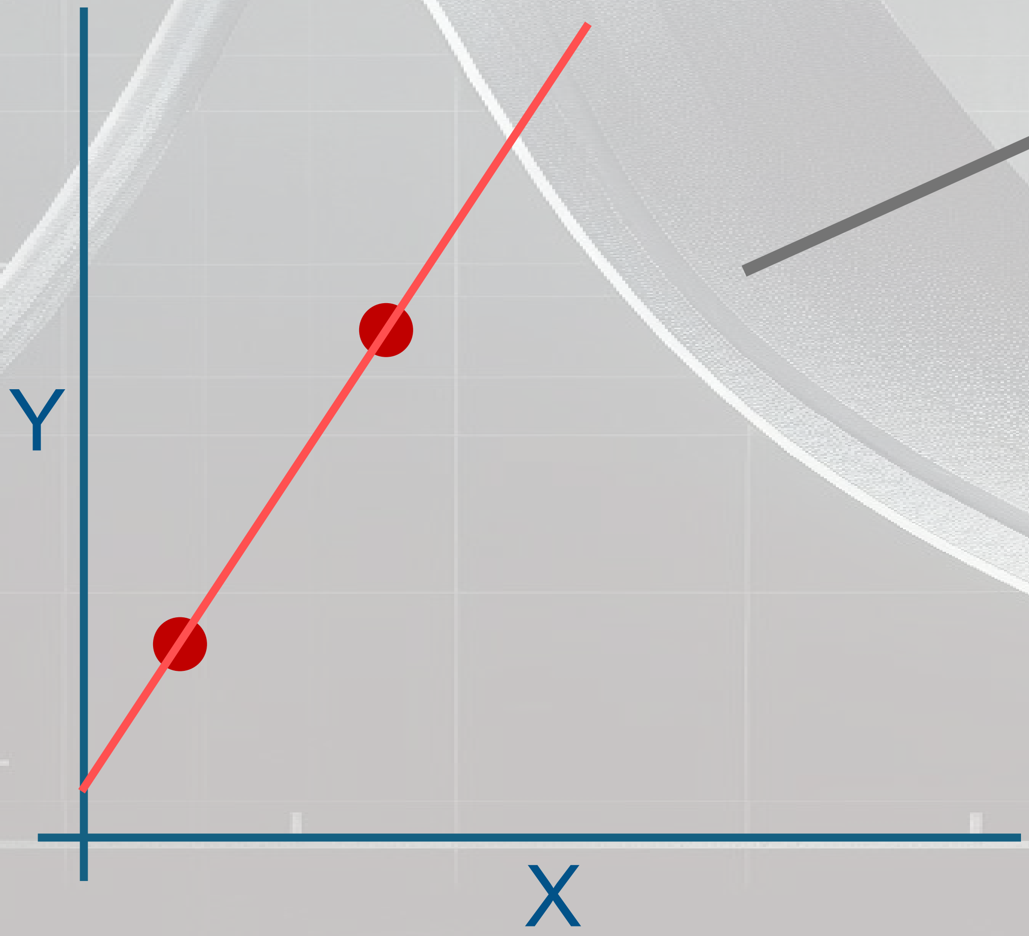
$$\text{OLS:} \quad \min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- Requires $\boldsymbol{X^T X}$ to be invertible.
- If we have a lot of observations, we can be we sure that the model reflects the relationship between X and Y.
- If predictors are correlated (multicollinearity), the matrix $\boldsymbol{X^T X}$ becomes near-singular:
    - inflating coefficient variance
    - making estimates unreliable

Introduction

But what if we only have two points?

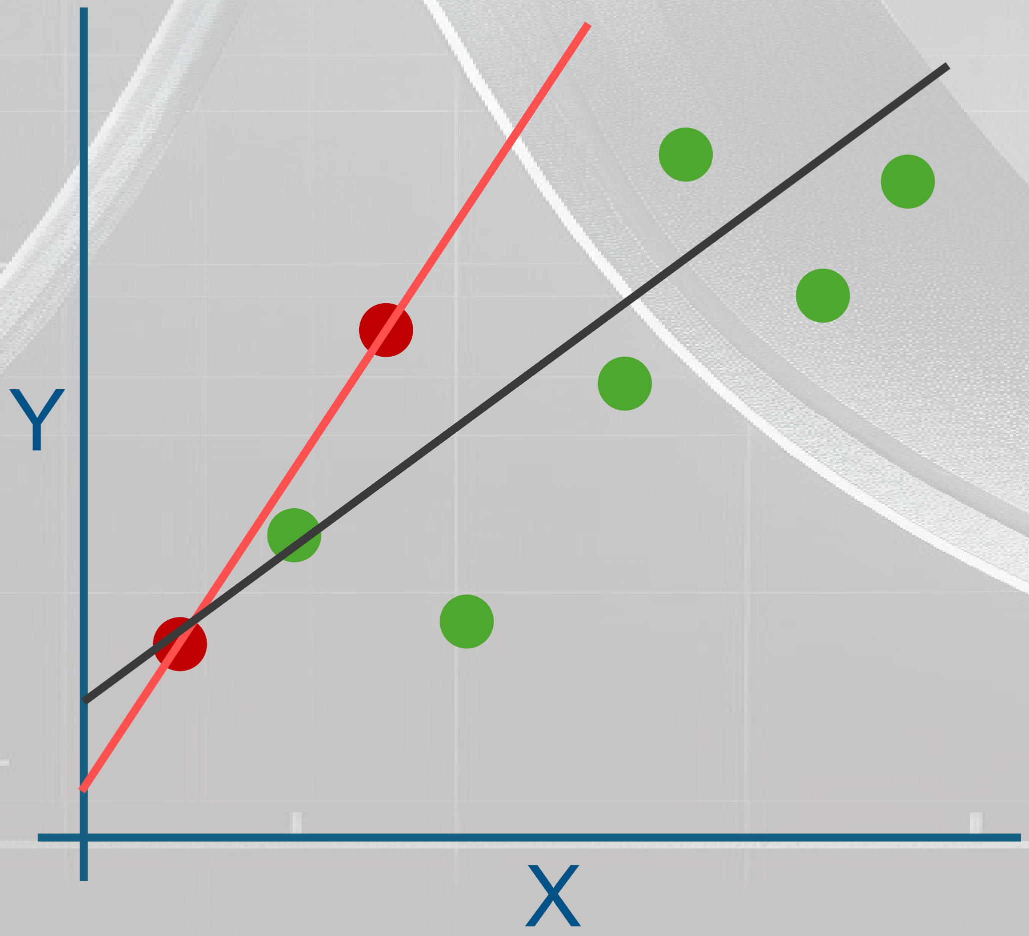


$$\min_{\beta} \sum_{i=1}^{2} (y_i - X_i\beta)^2 = 0$$

- If we have few points, the minimum sum of the residuals will be close to 0 because it is easier to find a model that fits well.

Introduction

Here are the original data and the original model for comparison.

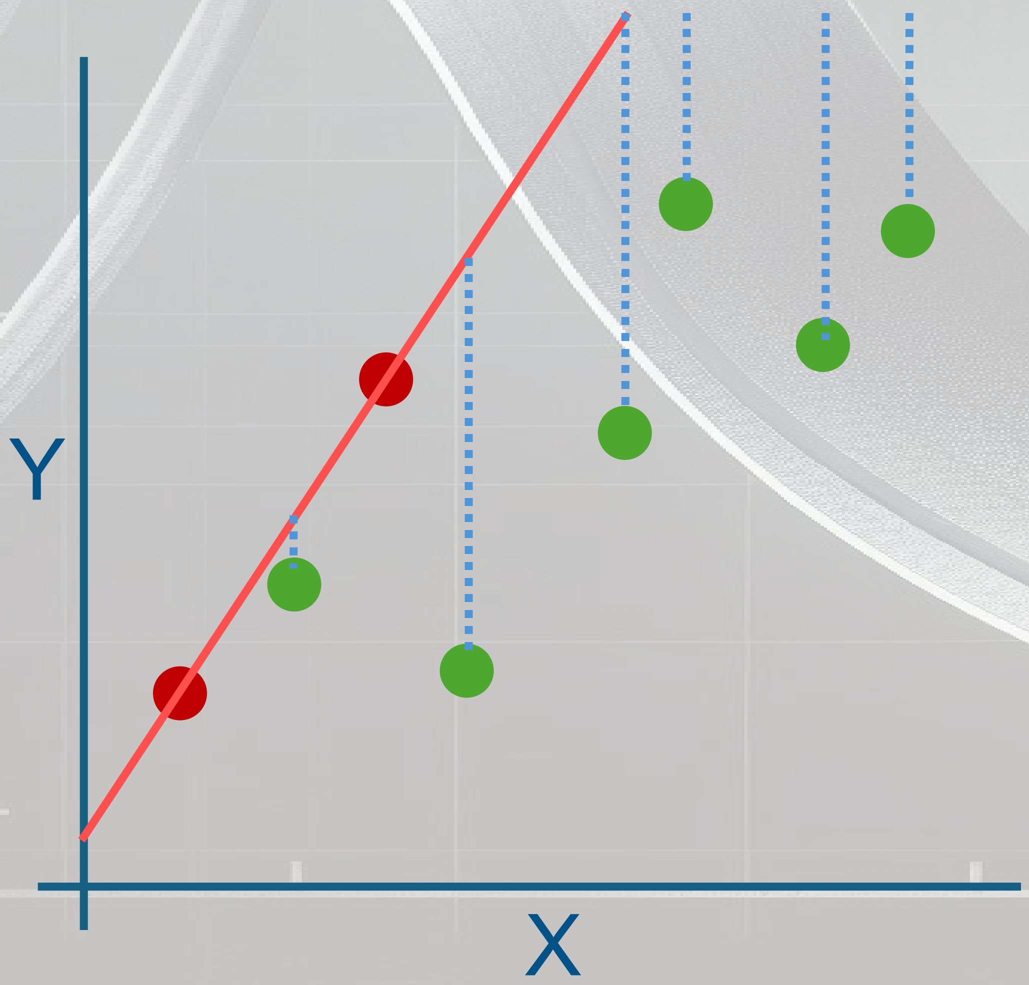


- ➢ In machine learning techniques we need to divide the dataset into two subsets: **training** and **testing**.

- ❖ Let's call the red dots the training data, and the remaining green dots the testing data.

Introduction

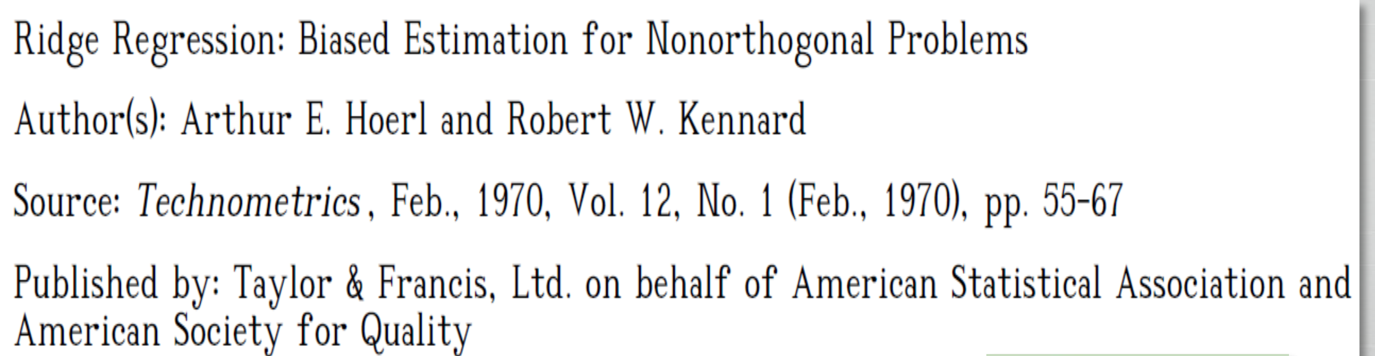
The sum of the squared residuals for the training data is small (0 in this case), but for the testing data is large.



- The red model has **high variance**.
- In machine learning, we'd say that the red model is **Over-Fit** to the training data.
- What if we introduce a small amount of bias into the red model?

Ridge regression

Hoerl and Kennard (1970)



Ridge Regression: Biased Estimation for Nonorthogonal Problems

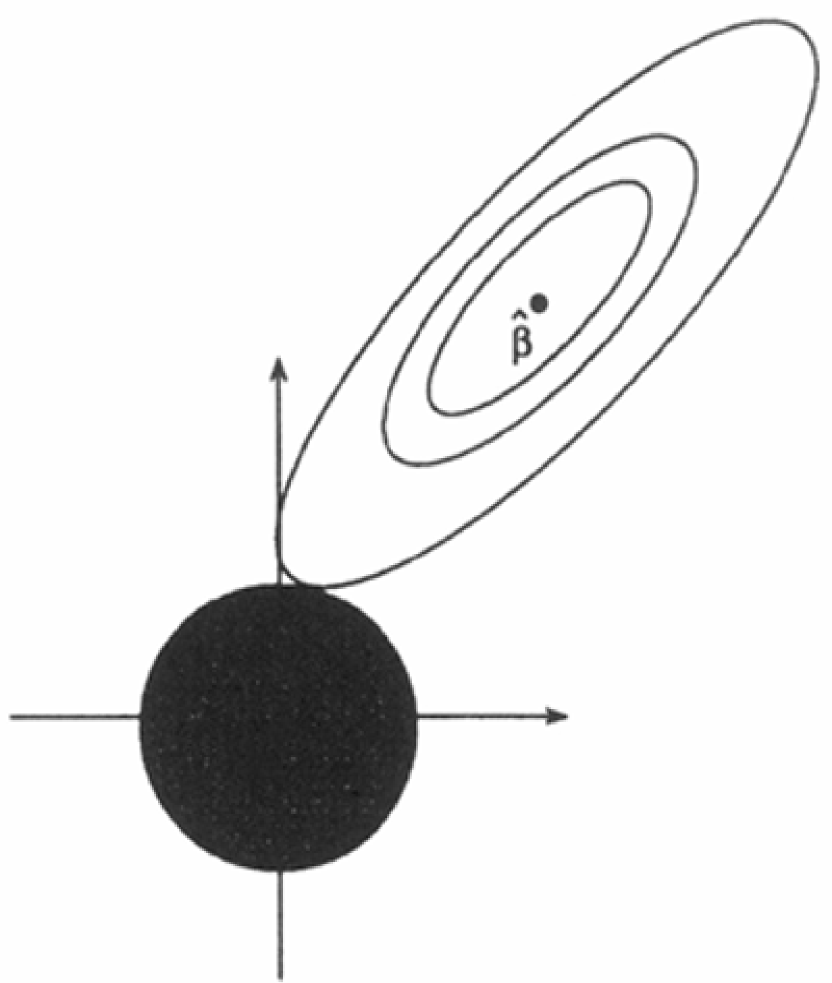
Author(s): Arthur E. Hoerl and Robert W. Kennard

Source: *Technometrics*, Feb., 1970, Vol. 12, No. 1 (Feb., 1970), pp. 55-67

Published by: Taylor & Francis, Ltd. on behalf of American Statistical Association and American Society for Quality

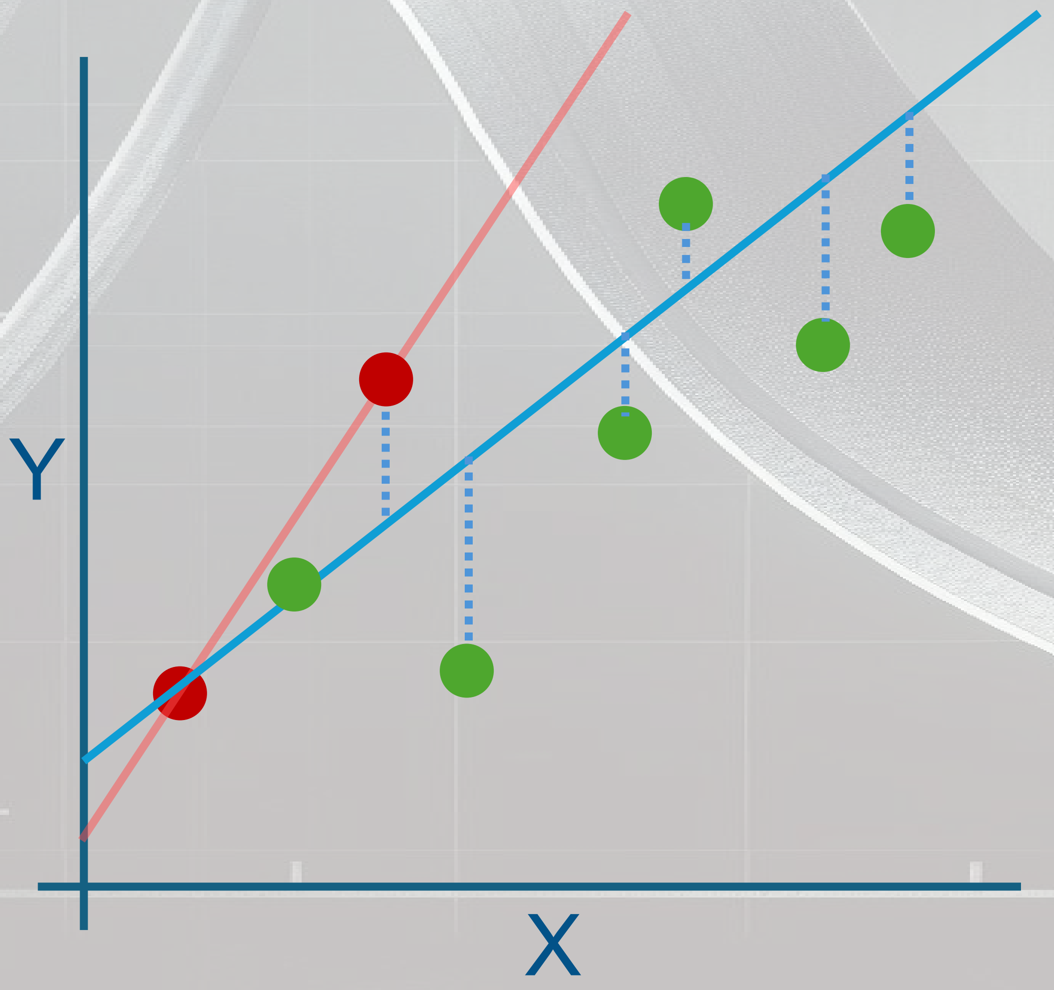
Stable URL: https://www.jstor.org/stable/1267351

Technometrics
A Journal of Statistics for the Physical, Chemical, and Engineering Sciences



Ridge regression

The idea behind **Ridge Regression** is to find a new model that doesn't fit the training data as well…



- …we introduce a small amount of **bias** in the way the model fits the data.
- But for that small amount of **bias**, we get a significant reduction in **variance**.
- That is, by starting with a worse fit, Ridge regression can provide better long-term predictions.

Ridge regression

Ridge adds a L2 penalty (sum of squared **slopes** *β_j*) to the OLS loss function:

$$\hat{\boldsymbol{\beta}}^{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

is equivalent to

$$\hat{\boldsymbol{\beta}}^{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

or

$$\hat{\beta}^{\text{Ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

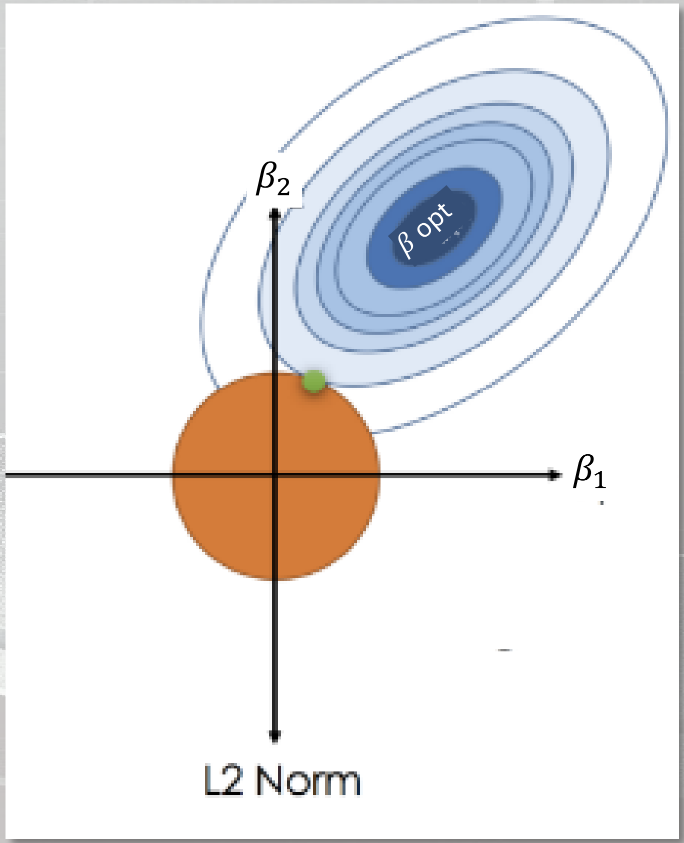
- $\lambda \geq 0$, the regularization parameter (controls penalty strength).
- $\sum \beta_j^2$: L2 norm of coefficients (excluding intercept β_0).
- *Stabilizes* $\boldsymbol{X^T X}$ by adding $\boldsymbol{\lambda I}$ (identity matrix) to OLS solution. This ensures invertibility even with multicollinearity.

Ridge regression

➢ Ridge regression pulls *β's* toward zero to minimize the new loss function.

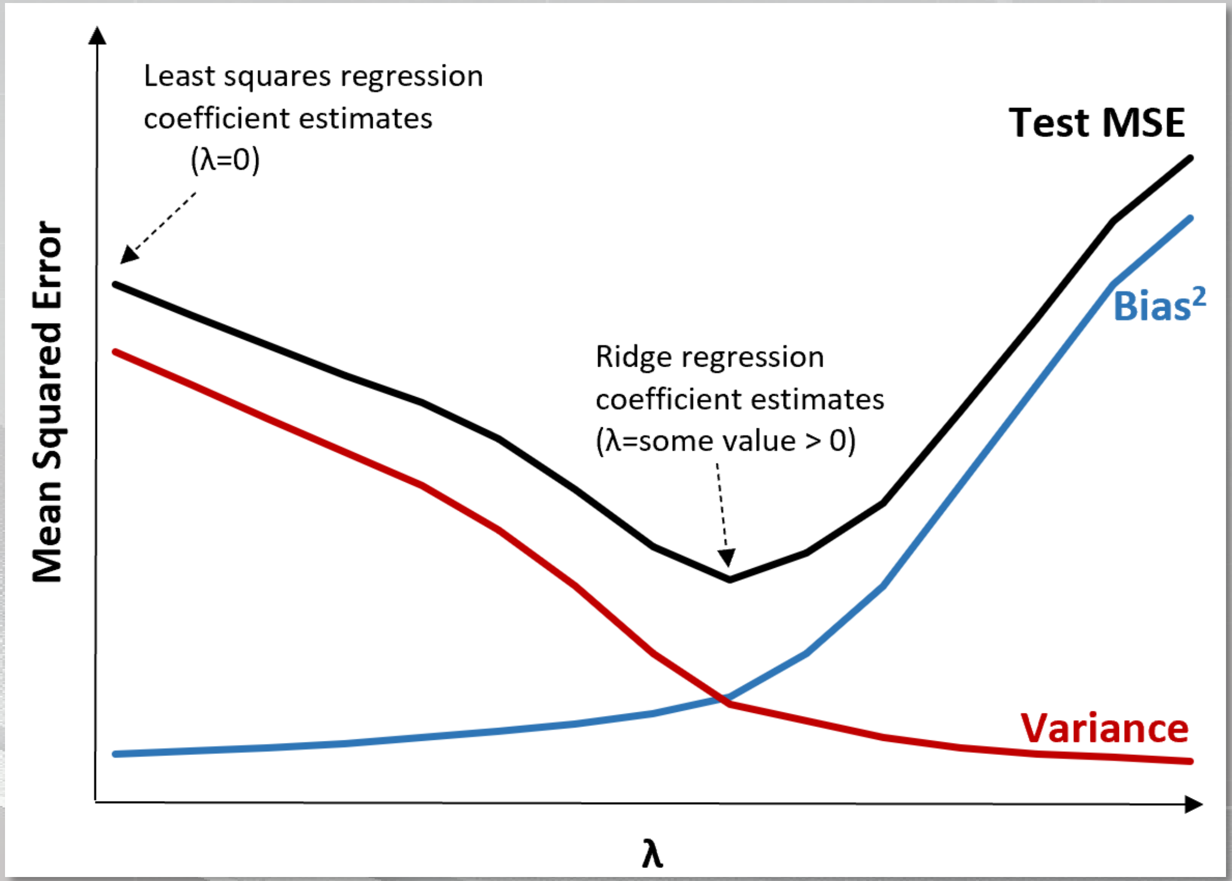
$$\underbrace{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}_{\text{OLS loss}} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{\text{L2 penalty}}$$

- **OLS**: Finds coefficients where the residual sum of squares (RSS) is minimized (unconstrained).
- **Ridge**: Constrains coefficients to lie within a **hypersphere** (L2 ball) centered at zero.
  - The solution is the point where the RSS contours touch the L2 ball tangentially.
  - The larger λ, the smaller the L2 ball, forcing coefficients (slopes) toward zero (but never zero).



Ridge regression

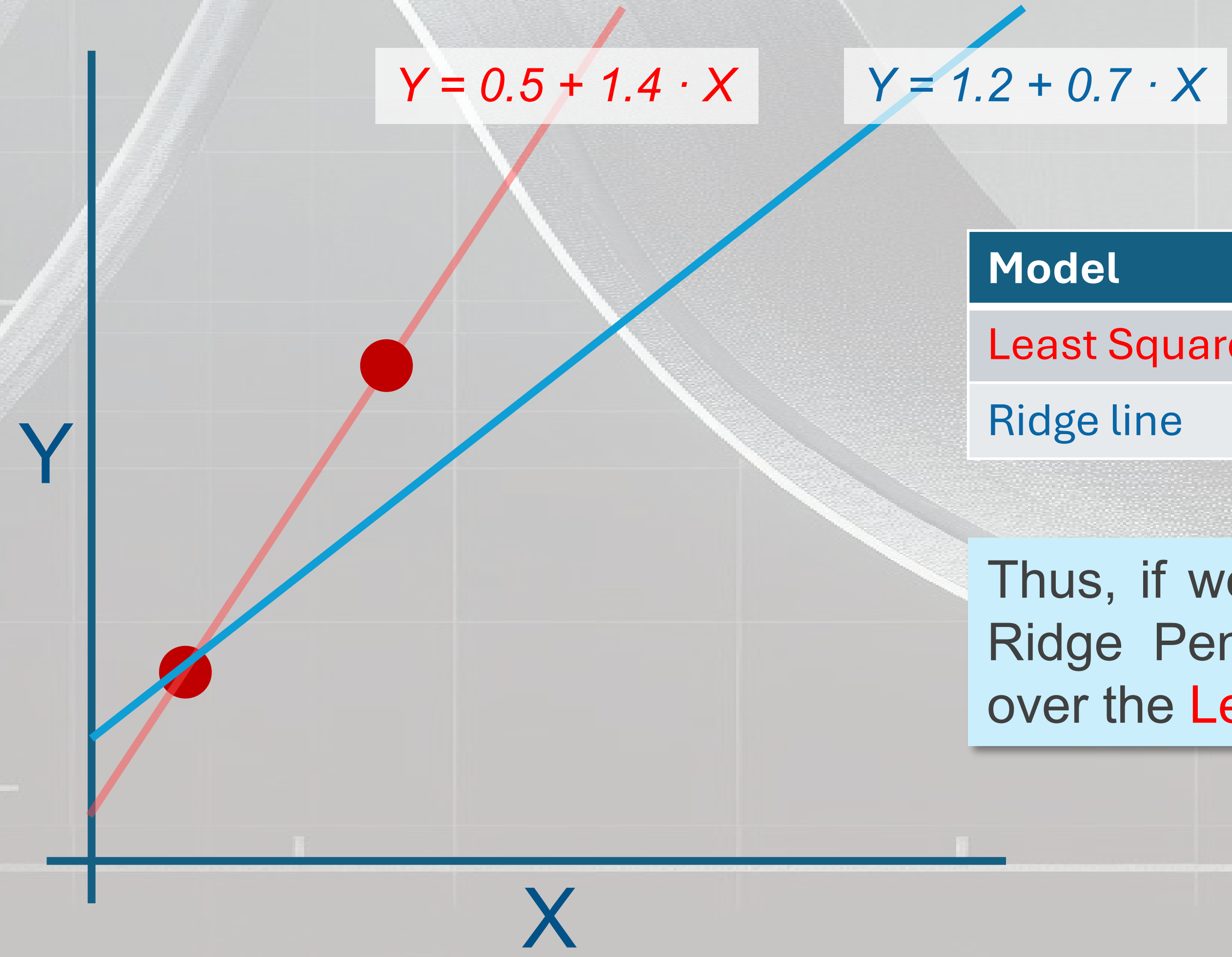
- *Bias-Variance tradeoff*:
  - ***Bias***: how much your model's predictions deviate from training data. Bias ↑ as *λ* increases.
  - ***Variance***: how much your model's predictions from the test data. Variance ↓ as *λ* increases.



- Predictors ***X*** must be standardized because penalization is scale-sensitive.
- λ is estimated using cross-validation.

Ridge regression

The effect of λ → Example, let λ = 1.

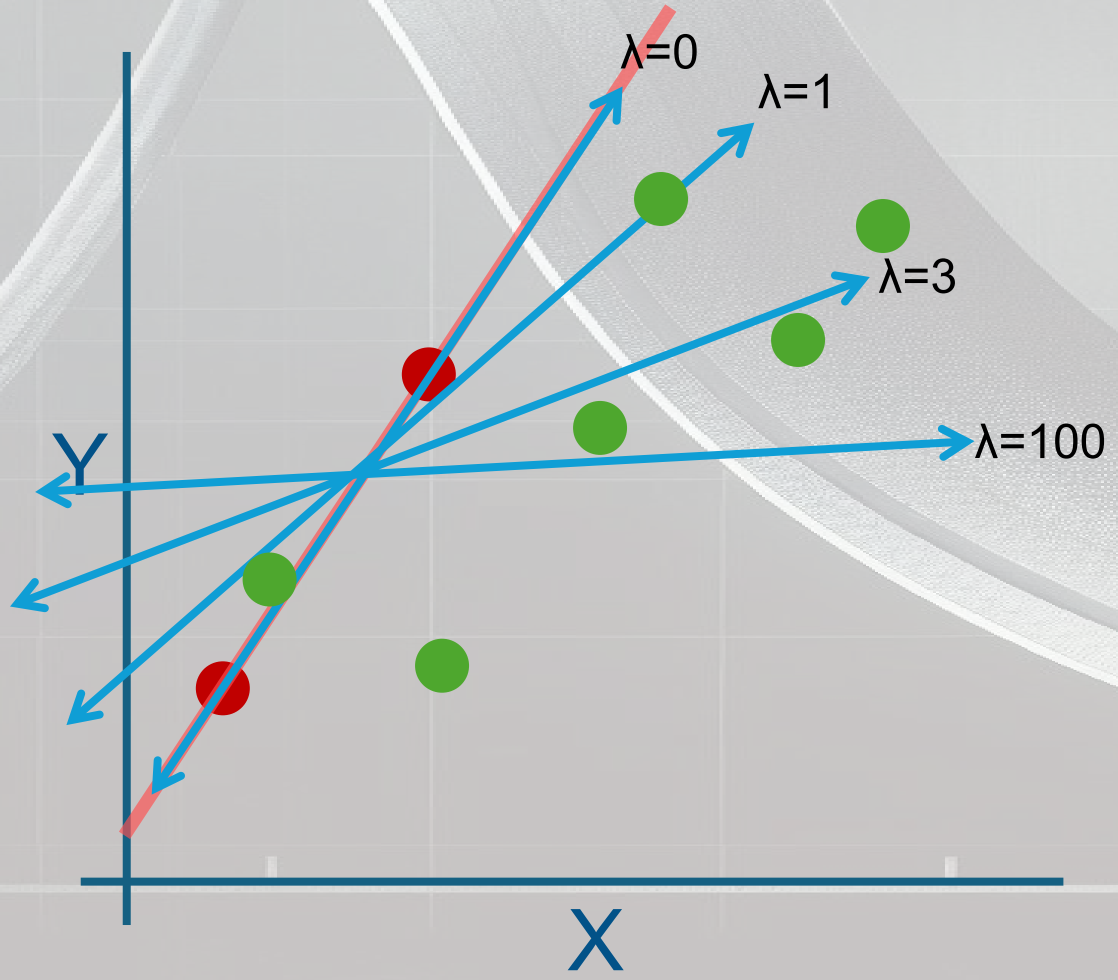


Model	SSR	λ x slope^2	Loss
Least Squares line	$0^2 + 0^2 = 0$	1 x 1.4^2	1.96
Ridge line	$0.1^2+1.1^2 = 1.22$	1 x $0,7^2$	1.71

Thus, if we wanted to minimize the SSR plus the Ridge Penalty, we would choose the Ridge line over the Least Squares line.

Ridge regression

The effect of λ → If we increase λ, the slope gets smaller to minimize the total loss function.



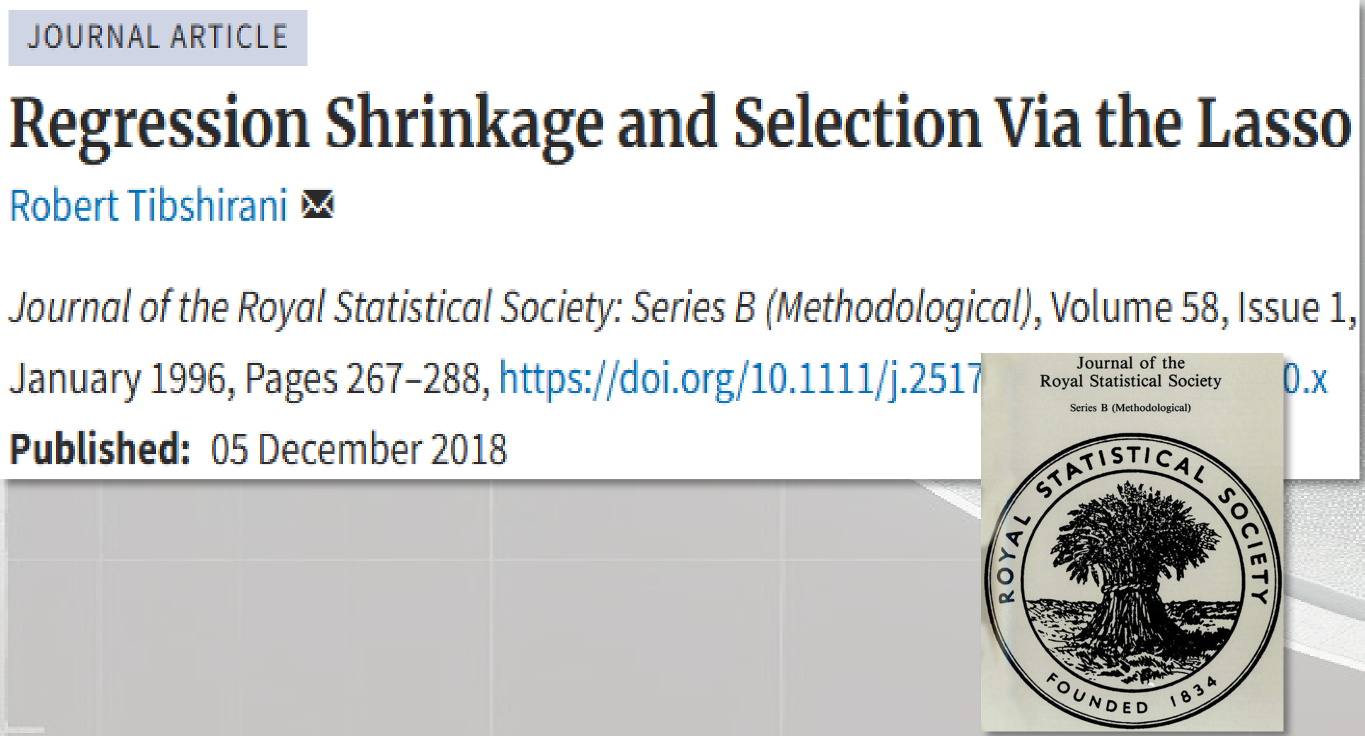
The larger is λ:

- ➢ slope tends asymptotically to 0.
- ➢ Y becomes less sensitive to X.

❑ Cross Validation (typically 10-fold) is used to determine the value of λ giving the best bias-variance

Lasso regression

Tibshirani R (1996)



JOURNAL ARTICLE

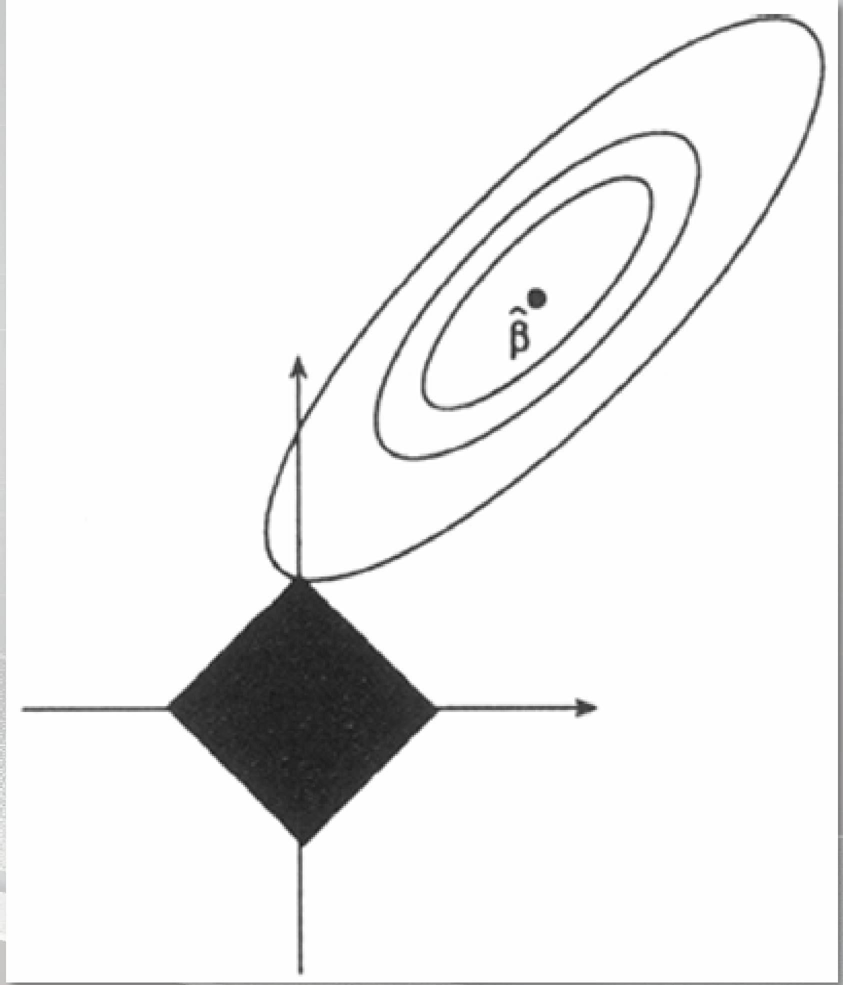
Regression Shrinkage and Selection Via the Lasso

Robert Tibshirani ✉

Journal of the Royal Statistical Society: Series B (Methodological), Volume 58, Issue 1, January 1996, Pages 267–288, https://doi.org/10.1111/j.2517 0.x

Published: 05 December 2018

Journal of the
Royal Statistical Society
Series B (Methodological)
ROYAL STATISTICAL SOCIETY
FOUNDED 1834



Lasso regression

Lasso adds a L1 penalty (sum of absolutes values of **slopes** β_j) to the OLS loss function. It works similarly to Ridge by changing the L2 norm to L1 norm.

$$\hat{\boldsymbol{\beta}}^{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

is equivalent to

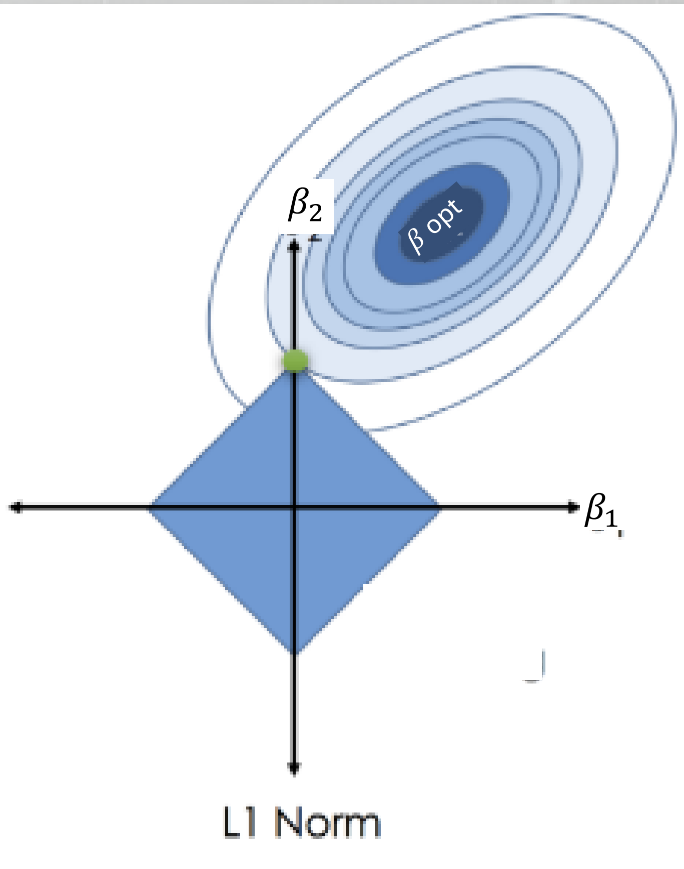
$$\hat{\boldsymbol{\beta}}^{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- $\lambda \geq 0$, the regularization parameter, controls penalty strength as in Ridge.
- Sum of L1 norm of coefficients $\sum|\beta_j|$ forces some β_j to be **exactly 0**.

Lasso regression

- *Effect on coefficients*: Ridge Regression produces a smooth shrinkage (no exact zeros), but Lasso selects variables (exact zeros).

$$\underbrace{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}_{\text{OLS loss}} + \lambda\underbrace{\sum_{j=1}^{p}|\beta_j|}_{\text{L1 penalty}}$$



L1 Norm

- **Lasso**: Constrains coefficients to lie within a **diamond** (2D) or a high-dimensional **polytope** centered at zero.
  - The solution is the point where the RSS contours touch the L1 diamond tangentially.
  - The larger λ, the smaller the L1 diamond, forcing coefficients (slopes) toward 0 or even to take the value 0.

Lasso regression

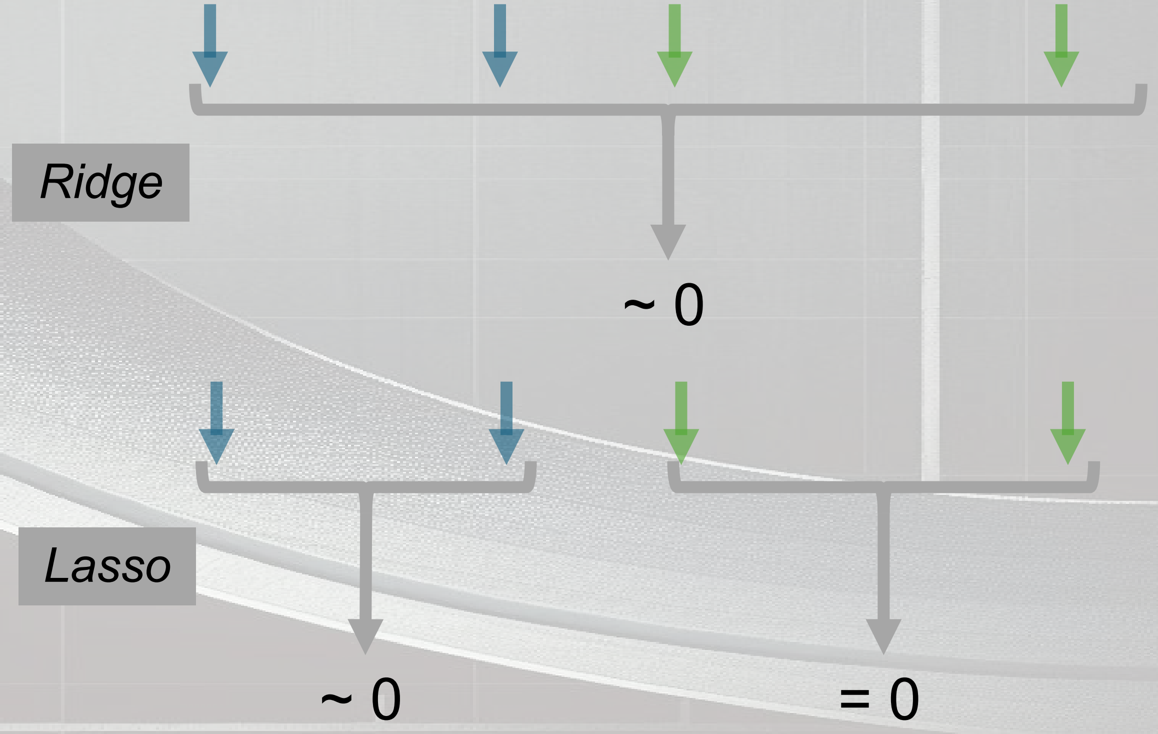
- Standardization required and choosing λ through cross-validation.

- *When to use Lasso?*
  - **Variable selection**: when you suspect that many characteristics are irrelevant.
  - **Interpretable models**: to obtain models with fewer predictors.
  - **High dimensional data**: if the number of predictors (p) is much larger than the number of samples (n).

idea: when λ increases:

relevant predictors *non-relevant predictors*

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p$$



Ridge

~ 0

Lasso

~ 0

= 0

Ridge vs Lasso

	Ridge Regression	Lasso Regression
Penalty type	$\lambda \sum_j \beta_j^2$	$\lambda \sum_j \lvert \beta_j \rvert$
Correlated Predictors	Similar weights to correlated predictors.	Selects one predictor and discards others.
Advantages & Disadvantages	Stable with multicollinearity. Good performance when *p*>*n*. No dimensionality reduction. Less interpretable for large *p*.	Automatic predictor selection. Interpretability (simpler models). Unstable with highly correlated predictors. May select only *n* predictors if *p*>*n*.
Typical Use Case	High multicollinearity. All predictors are relevant.	Removing irrelevant predictors. Interpretable models.

Elastic-net regression

Zou & Hastie (2005)

JOURNAL ARTICLE

Regularization and Variable Selection Via the Elastic Net

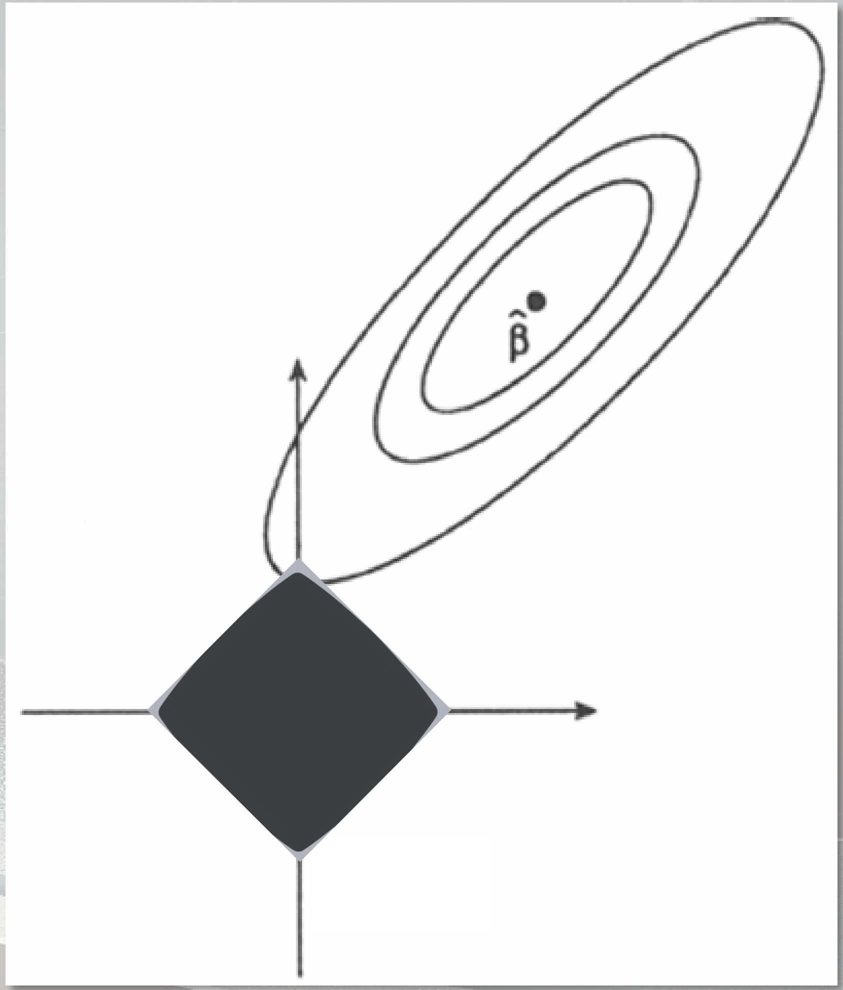
Get access ›

Hui Zou , Trevor Hastie ✉

Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 67, Issue 2, April 2005, Pages 301–320, https://doi.org/10.1111/j.1467-986

Published: 09 March 2005 **Article history** ▾

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Series B
Statistical
Methodology
VOLUME 87 PART 2
B
APRIL 2025



Elastic-net regression

Elastic-net is a regularized regression method that combines the Lasso and Ridge penalties to overcome limitations when there are more predictors than observations or when there are highly correlated variables.

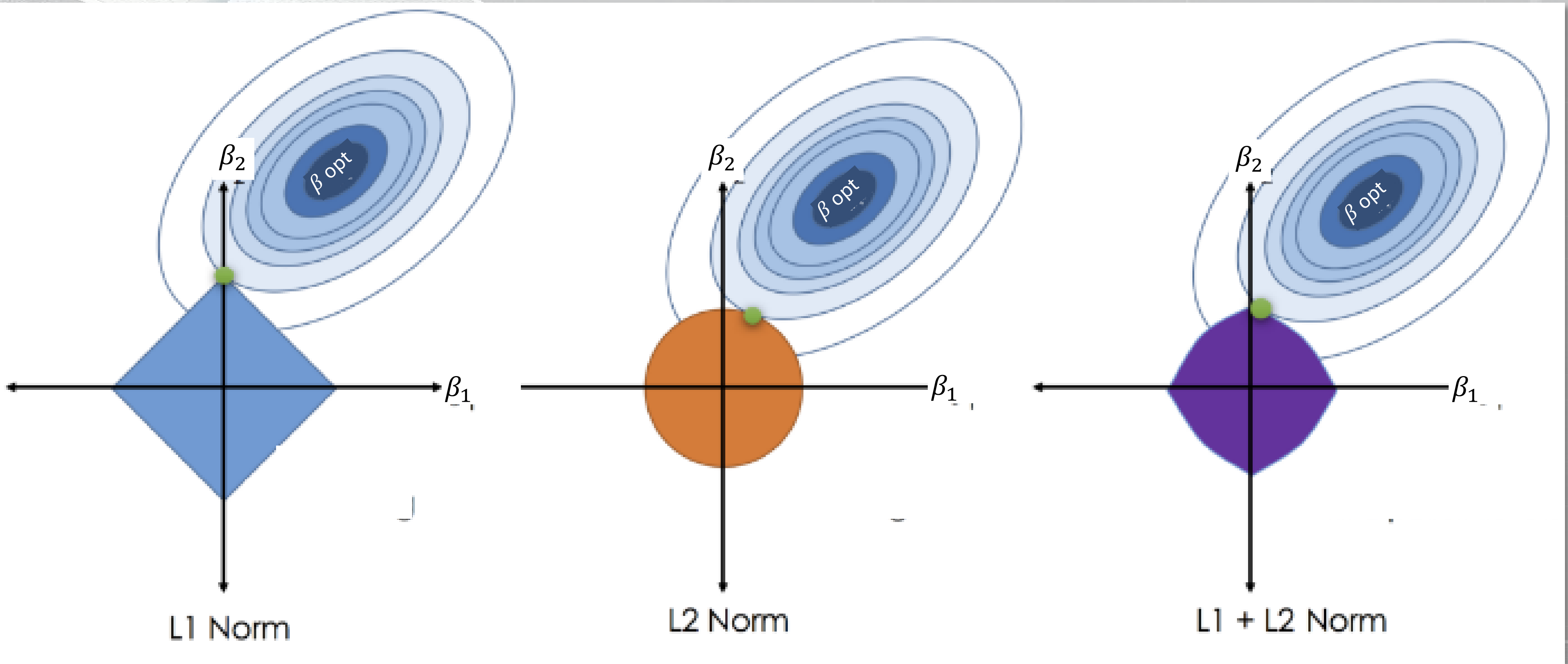
$$\hat{\beta}^{\text{Elastic-Net}} = \arg\min_{\beta}\left\{\|\mathbf{y}-\mathbf{X}\beta\|_2^2 + \lambda\left(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2\right)\right\}$$

is equivalent to

$$\hat{\beta}^{\text{Elastic-Net}} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2\right)\right\}$$

- ➢ $\lambda \geq 0$ → controls penalty strength.
- ➢ $0 \leq \alpha \leq 1$ → determines the mix between L1 y L2.
- ➢ $\alpha = 1$ → Lasso , $\alpha = 0$ → Ridge.

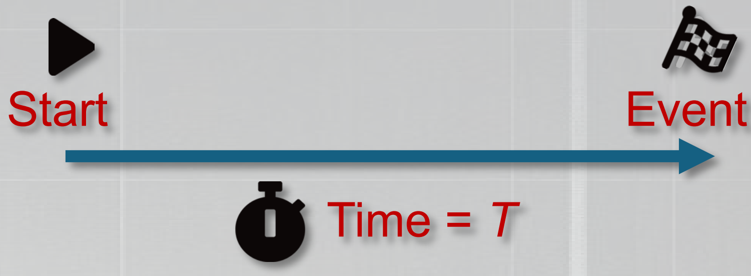
Regularization regression methods



	Lasso	Ridge	Elastic-Net
Variable selection	Yes	No	Yes (less aggressive)
Correlated predictors	Randomly selects one	Assigns similar weights	Group and select

Survival analysis

The branch of statistics focused on analyzing the time until an event occur (death, recurrence, failure…)



- **Survival function**: probability that the event occurs beyond a time *t*.

$$S(t) = P(T > t)$$

- **Hazard function**: the probability that if you survive to t, you will experiment the event in the next instant.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Goals:

1) Estimate survival function over time.
2) Compare survival between different groups of individuals.
3) Identify risk factors associated with survival and quantify their influence.

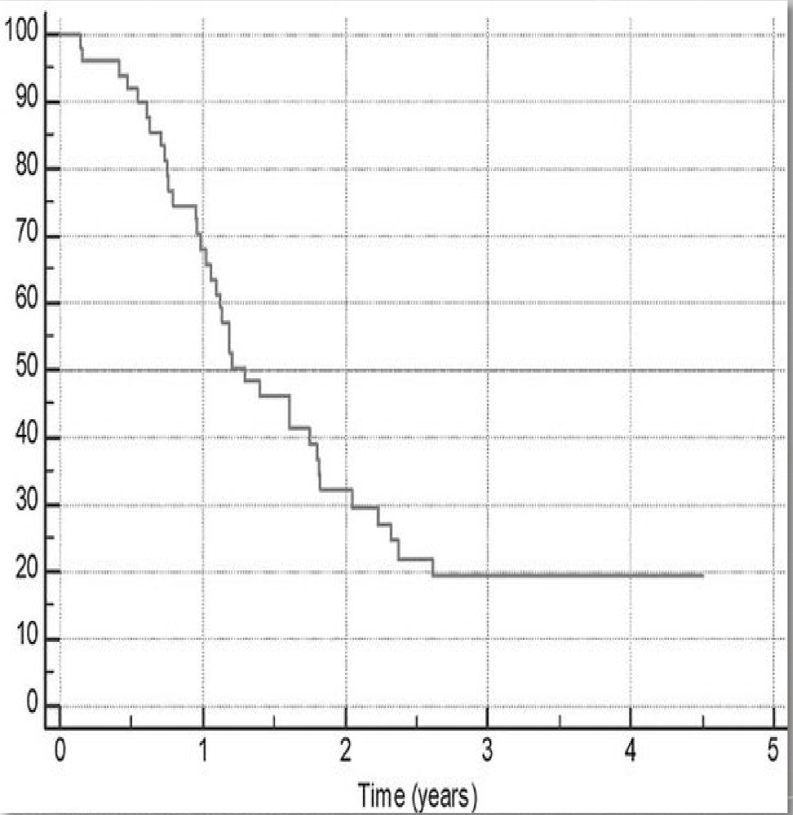
Survival analysis

- **(1) Survival function estimation**:

  Kaplan-Meier method (Kaplan, Meier 1958)

  - t_i: distinct event times (ordered, $t_1 < t_2 < \cdots t_n$)
  - d_i : number of events in t_i
  - n_i : number of individuals at risk just before t_i
  - $h_i = \frac{d_i}{n_i}$: risk of the event in $[t_i, t_{i+1})$

$$\hat{S}(t) = \prod_{t_i<t}(1 - h_i) = \prod_{t_i<t}\left(1 - \frac{d_i}{n_i}\right)$$



Survival analysis

- **(2) Survival comparation: Log-rank test (Mantel-Cox test)** (Mantel 1966)

  ***H*$_0$**: All groups have the same survival function: $S_1(t) = S_2(t) = \ldots = S_k(t)$

  ***H*$_1$**: At least one group differs in survival: $S_i(t) \neq S_j(t)$ *, for some* $i \neq j$

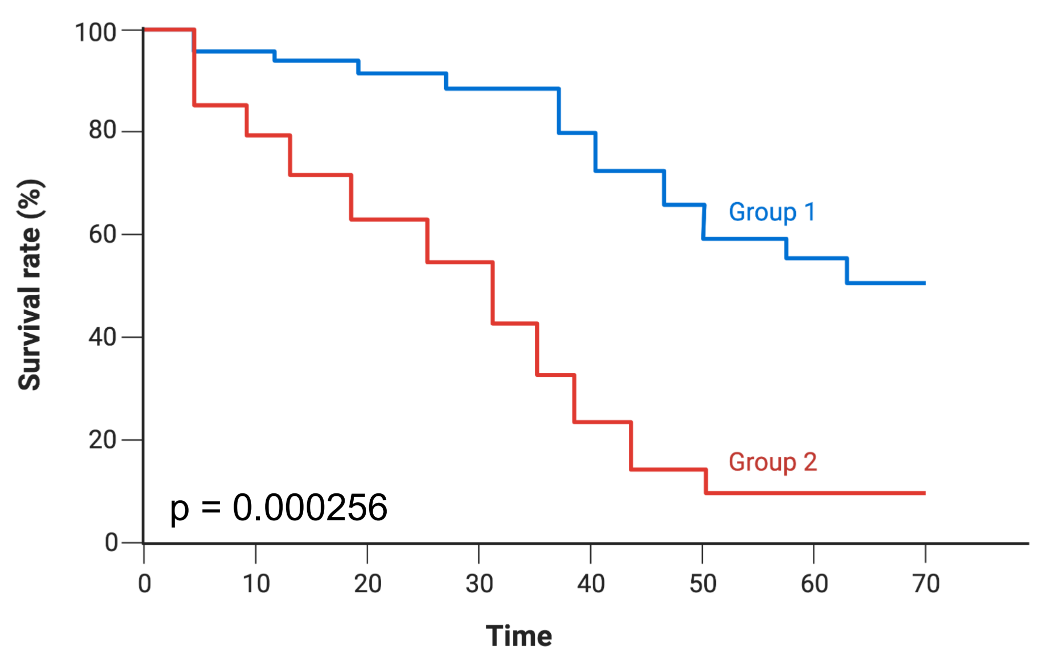
Test statistic *is based on observed and expected events (χ^2 Pearson or Mantel-Haenszel)*

***O*$_i$**: vector of observed events for each group

***E*$_i$**: vector of expected events under H_0

$$\boldsymbol{U} = \sum_{\boldsymbol{i}} (\boldsymbol{O_i} - \boldsymbol{E_i}) \quad ; \quad \mathbf{V} = \sum_{\boldsymbol{i}} \boldsymbol{V_i}$$

$$\boldsymbol{\chi^2_{k-1} \sim U^* V^{*-1} U^*} \qquad (U^* \text{ and } V^* \text{ exclude the k}^{\text{th}} \text{ group})$$



Survival analysis

- **(3) Influence of risk factors: Cox proportional hazards model** (Cox 1972)

$$h(t|x) = h_0(t) \cdot exp\{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\}$$

$h_0(t)$: baseline risk

β_k: effect of factor *k*

It estimates the influence of *p* factors (covariates) in the event happening.

Coefficients β_j are estimated by maximizing the partial likelihood:

$$L(\boldsymbol{\beta}) = \prod_{i:\text{evento}} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_j)}$$

$$Example: X = \{0,1\} \longrightarrow \frac{h(t|x=1)}{h(t|x=0)} = \frac{h_0(t) \cdot exp\{\beta \cdot 1\}}{h_0(t) \cdot exp\{\beta \cdot 0\}} = e^{\beta}$$

What are you getting at?

Applications

Analysis of disease SUrvival and patient RIsk prediction based on gene signatures

Doctoral Thesis

Exploration and development of bioinformatics methods for survival analysis and drug targeting in cancer

Alberto Berral González (december 2024)

R

Package ASURI

REVIEW

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Applications: gene-phenotype

- Target: discovery of gene markers by identification of the significant association of gene expression (or another gene-related activity signal) with clinical variables or phenotypic characteristics (*G = {0, 1}*)

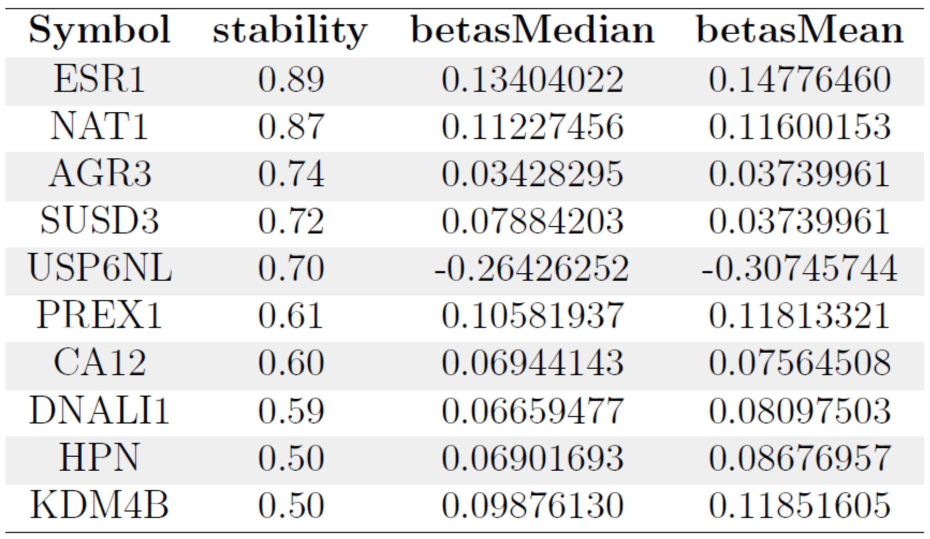
❖ Fit a classifier to the dataset based on bootstrapping and ensemble Elastic-Net models, (Friedman, 2010).

$$log\frac{Pr(G=1|x)}{Pr(G=0|x)} = \beta_0 + \beta^T x$$

$$P_\alpha(\beta) = \sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\cdot\beta_j^2 + \alpha\cdot|\beta_j|\right]$$

$$\max_{(\beta_0,\beta)\in\mathbb{R}^{p+1}}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(I(g_i=1)\cdot\log p_i + I(g_i=0)\cdot\log(1-p_i)\Big) - \lambda\cdot P_\alpha(\beta)\right]$$

(optimal regularized parameters using 10-fold CV)



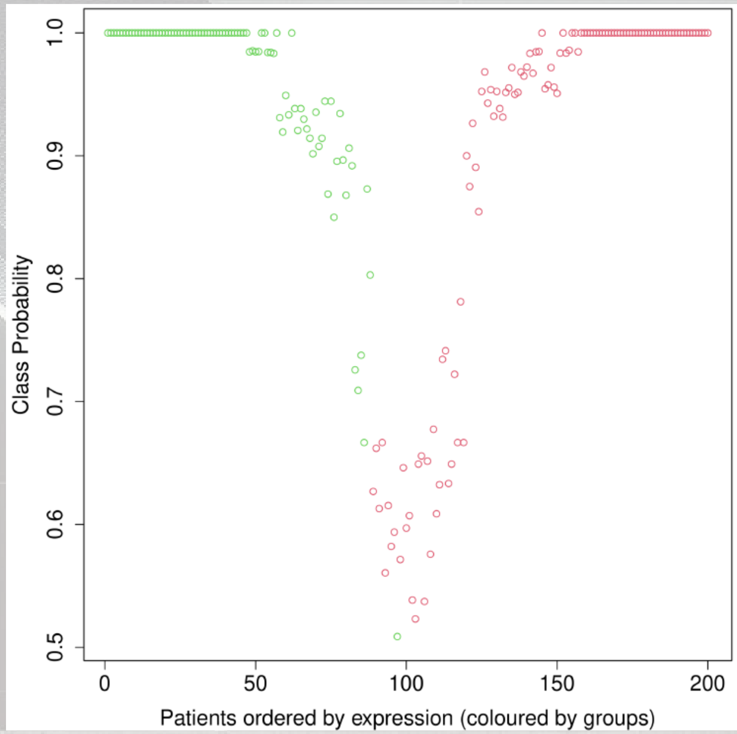
Symbol	stability	betasMedian	betasMean
ESR1	0.89	0.13404022	0.14776460
NAT1	0.87	0.11227456	0.11600153
AGR3	0.74	0.03428295	0.03739961
SUSD3	0.72	0.07884203	0.03739961
USP6NL	0.70	-0.26426252	-0.30745744
PREX1	0.61	0.10581937	0.11813321
CA12	0.60	0.06944143	0.07564508
DNALI1	0.59	0.06659477	0.08097503
HPN	0.50	0.06901693	0.08676957
KDM4B	0.50	0.09876130	0.11851605

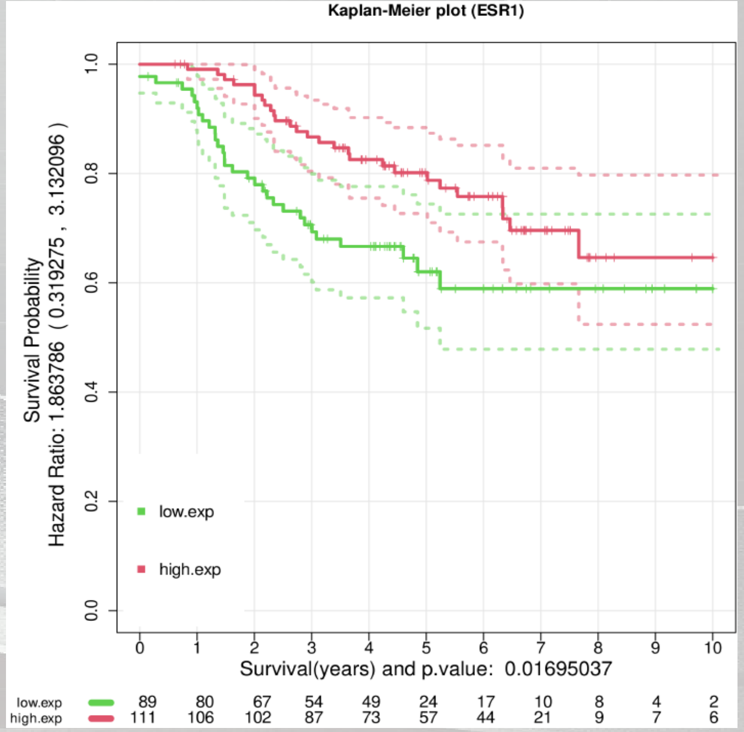
List of genes ordered by stability for the BRCA training dataset from Bueno-Fortes et al., 2023.

Applications: gene-survival

- Target: Discovery of robust and reproducible gene lists associated with disease survival based on gene expression (or another gene-related activity signal).

- Evaluate each gene as a prognostic marker by dividing patients into two groups (low/high expr.) with a threshold that we estimated by minimizing the *p-value* of the log-rank statistic.
- Strategy that determines the optimal p-value of the log-rank test that maximizes the separation of the Kaplan-Meier curves.





Applications: patient-risk

- Target: Construction of robust patient risk predictors based on gene signatures using univariate and multivariate Cox regression model approaches.

- Estimate patient risk with the Cox proportional hazards regression model but…

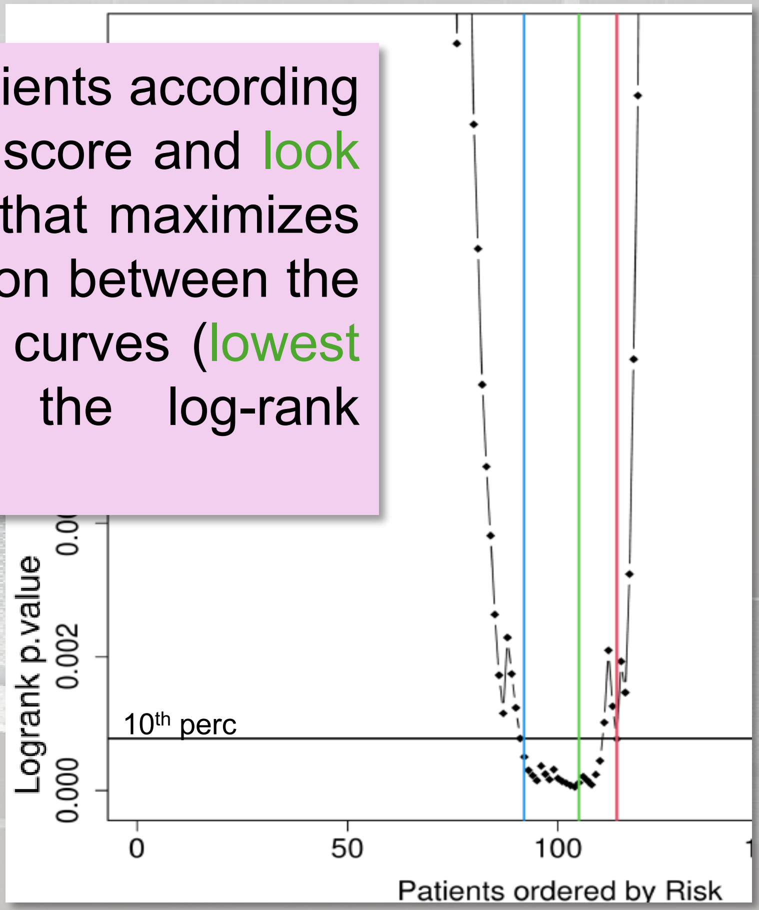
$$h(t|x) = h_0(t)\cdot exp\{\beta_1 X_1 + \cdots + \beta_p X_p\}$$

- …the β_j coefficients are estimated by maximizing the partial log-likelihood with a L1 (lasso) norm penalty. (Tibshirani, 2009).

$$l(\beta) = \sum_{j=1}^{p}\sum_{k=1}^{n}\left(x_{kj}\beta_j - \log\sum_{m\in\mathcal{R}_k}\exp(x_{mj}\beta_j)\right) - \lambda\cdot\sum_{j=1}^{p}|\beta_j|$$

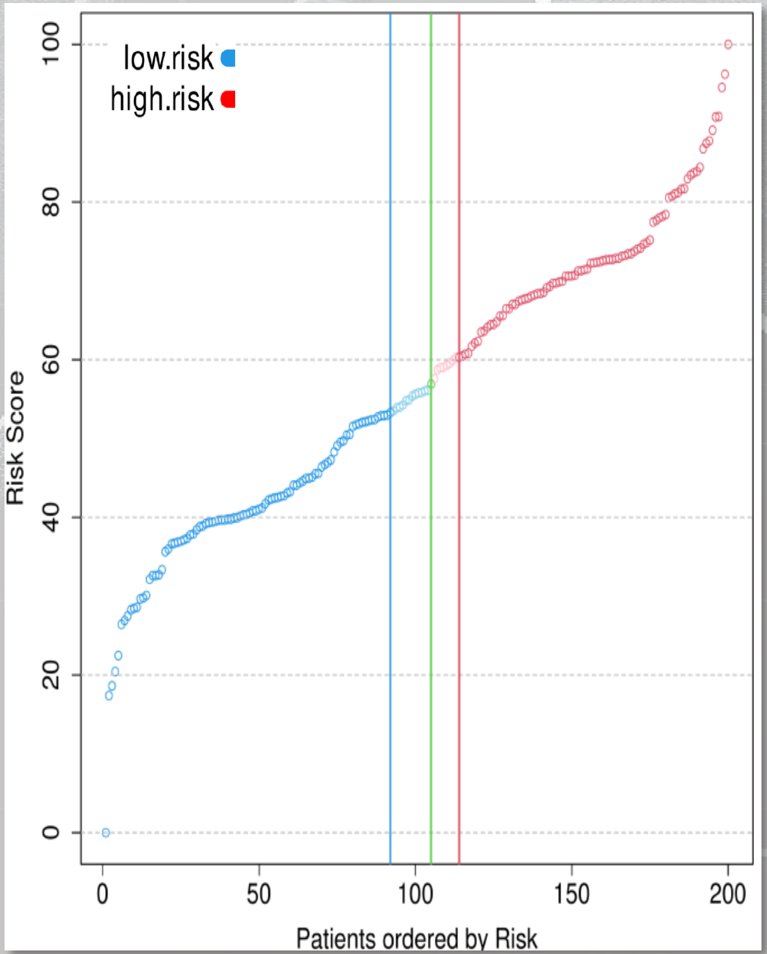
(optimal regularized parameters using 10-fold CV)

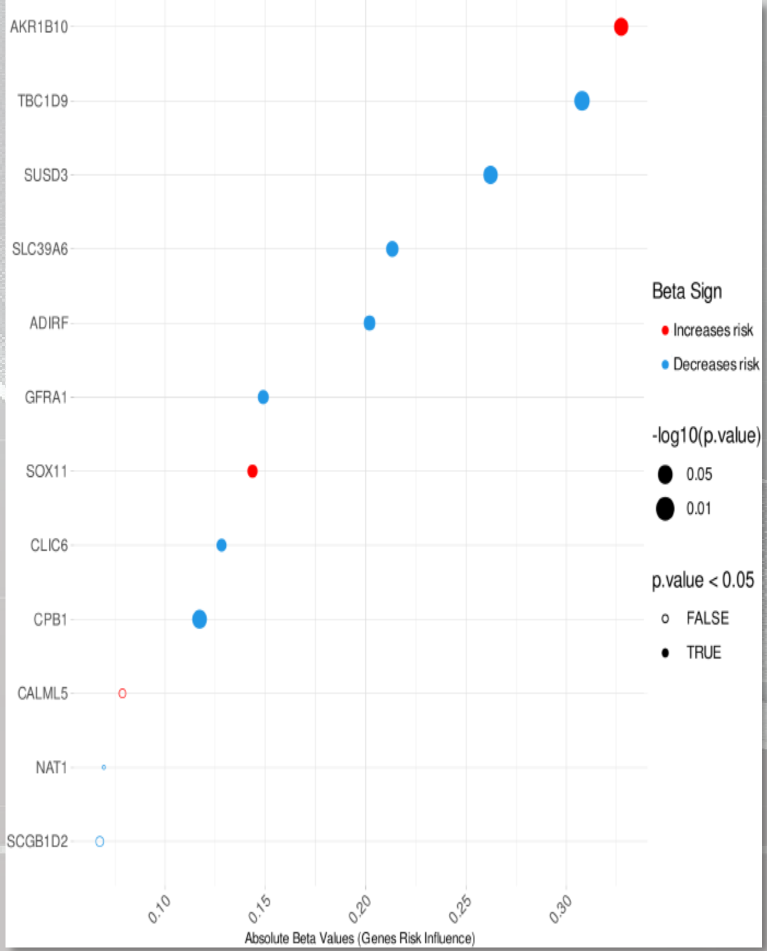
- We rank patients according to their risk score and look for the one that maximizes the separation between the KM survival curves (lowest p-value of the log-rank test).

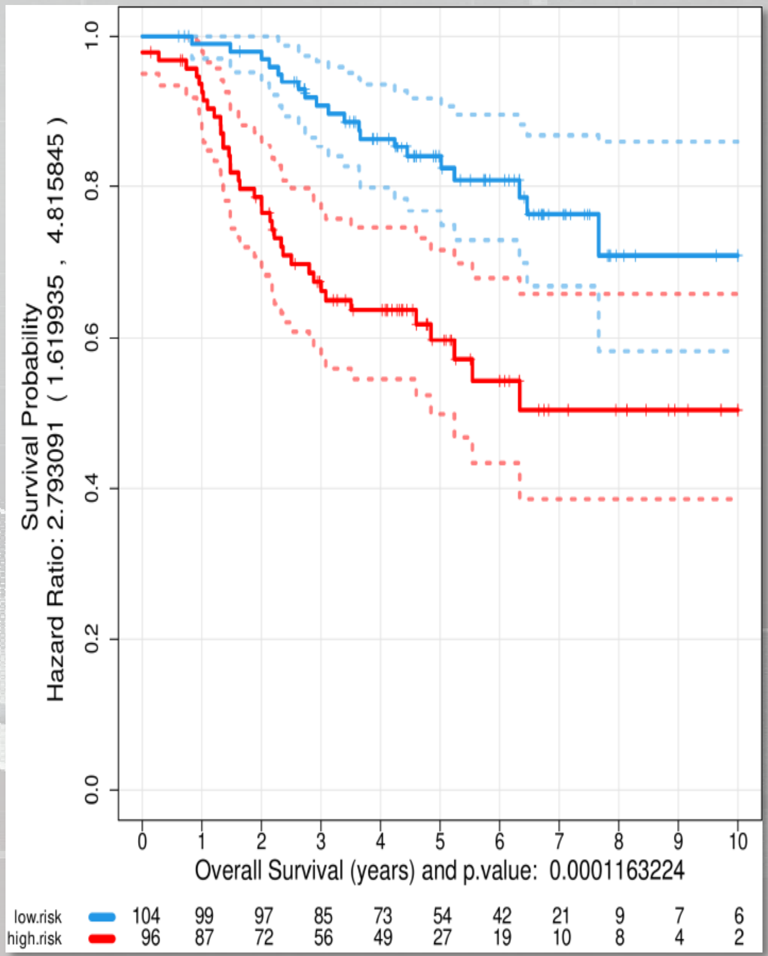


Applications: patient-risk

- The threshold allow us to separate the patients into two risk groups, low/high (*or three, in case we want to consider intermediate risk*).







Application to Breast cancer (BRCA)

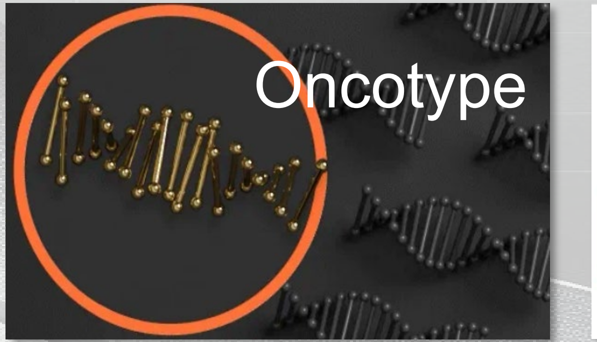
The predictive **IHC** (*immunohistochemistry*) markers in breast pathology include two cell proliferation markers and three hormone receptor positive factors (and their genes):

Chromosome segregation mitosis: **AURKA** / DNA damage: **MKI67**

Estrogen receptors: **ER** *(ESR1 gene)* / Progesterone receptors: **PR** *(PGR gene)*

Human epidermal growth factor receptor-2: **HER2** *(ERBB2 gene)*

❑ Two of the most widely used commercial platforms (Oncotype and Prosigna) use their own gene signatures to predict risk and stratify patients.



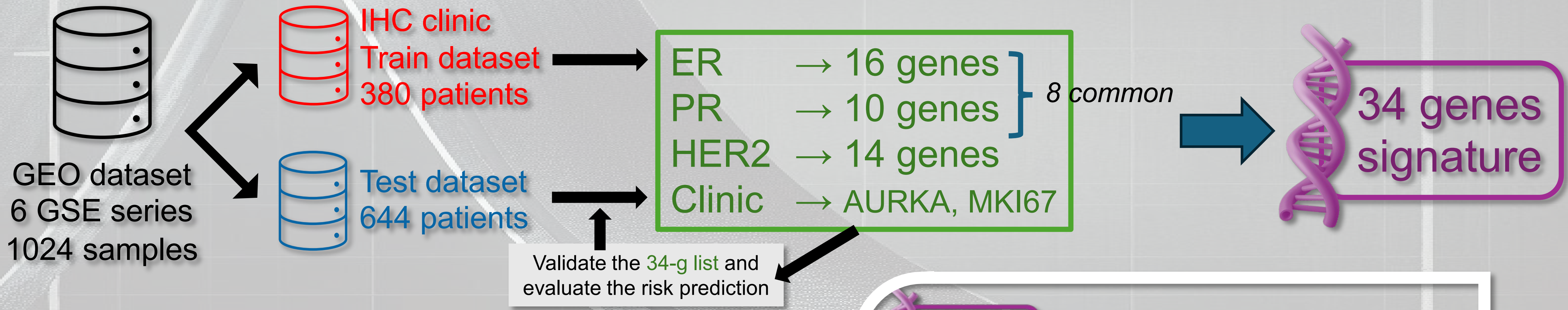
16 genes

50 genes

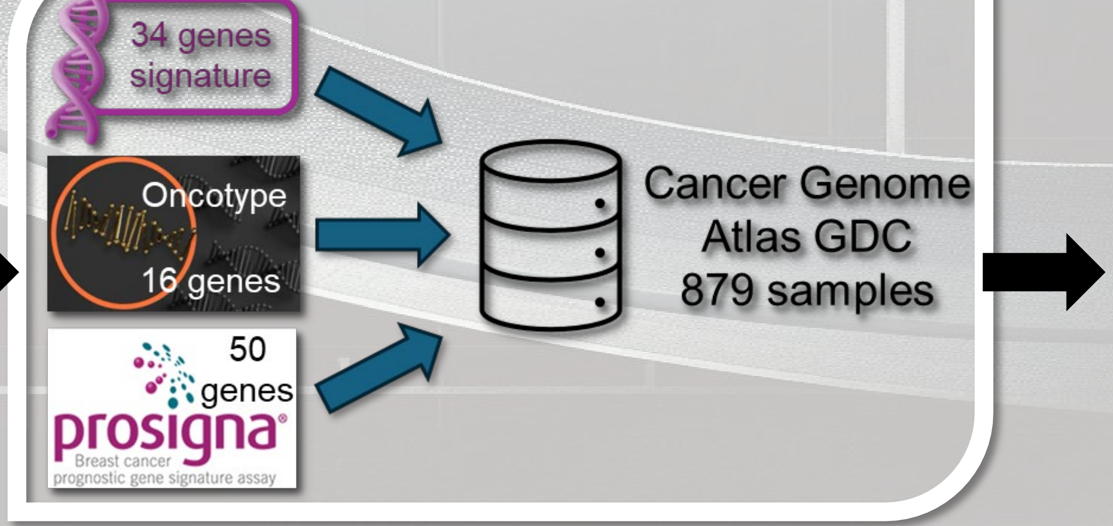
➢ Our goal is to identify survival markers related with that improve risk prediction and patient stratification better than these two

Application to Breast cancer (BRCA)

- ➢ We follow the approach described before and apply it to two independent BRCA datasets that integrate multiple primary tumor samples (*curated*, Bueno-Fortes, 2023)

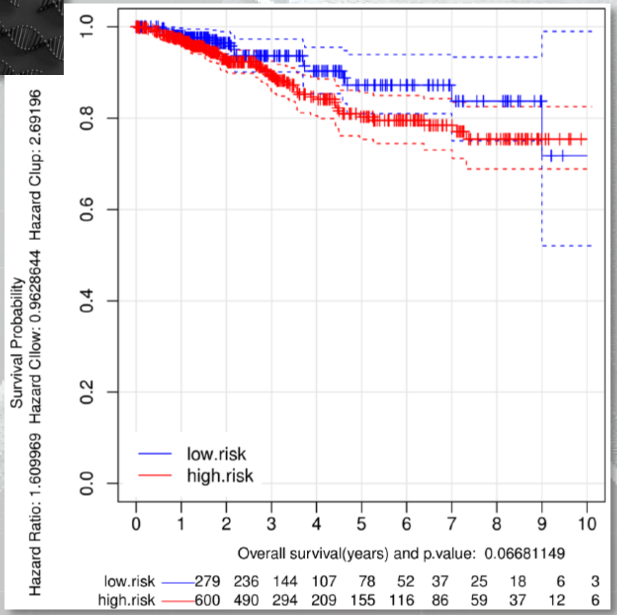


✓ We validated the prognostic power in a second dataset of 879 tumor samples and performed risk assignments for each sample.

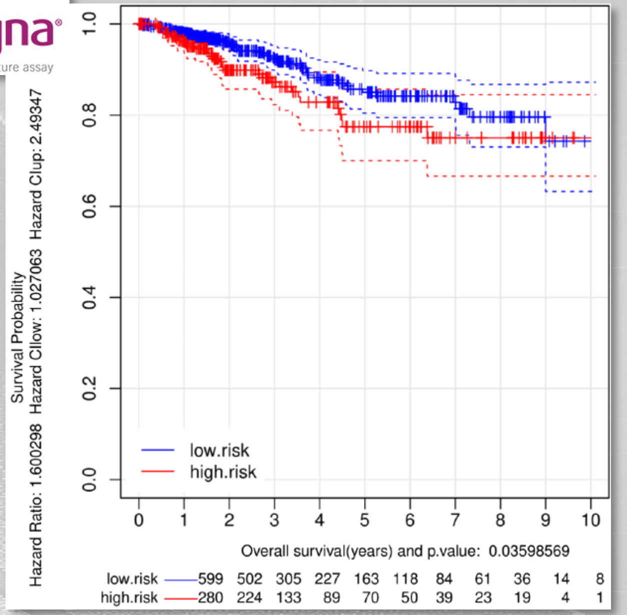


Application to Breast cancer (BRCA)

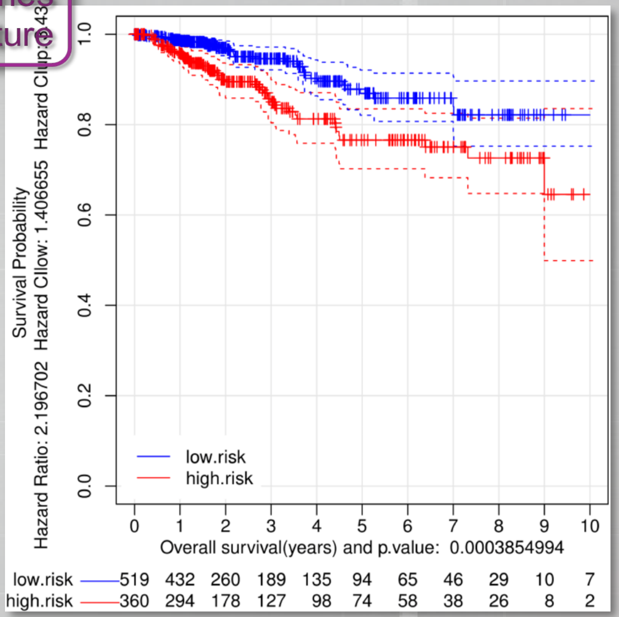
Oncotype



prosigna®
Breast cancer
prognostic gene signature assay



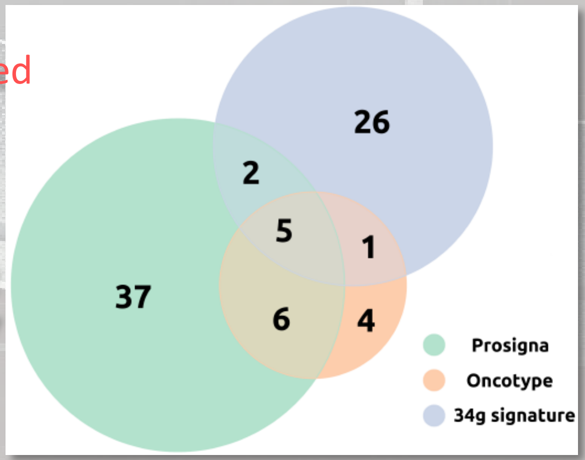
34 genes
signature



Signature	Log-rank p-value	Hazar ratio HR	95%CI of HR
34-g signature	0.00038	2.20	1.41 – 3.43
Oncotype 16-g	0.066	1.61	0.96 – 2.69
Prosigna 50-g	0.035	1.60	1.03 – 2.49

5 shared genes

ESR1
PGR
ERBB2
MKI67
GRB7



Some conclusions

- Techniques such as Elastic-net or Lasso ensure diversity and reliability to obtain robust survival and risk markers.

- The use of univariate or multivariate Cox regression and cross-validation leads to better selection of stable risk markers and better stratification of patients.

- We have applied a survival analysis methods for large human cancer datasets to validate previously established biomarkers and discover new ones with potential clinical relevance.

References

- Alfonsín, G., Berral-González, A., … (2024). Stratification of colorectal… IJMS, 25, 1919. 10.3390/IJMS25031919
- Berral-González, A. PhD thesis (2024). Exploration and development of bioinformatics methods for survival analysis… 10.14201/gredos.163622
- Bueno-Fortes, S., …, Berral-Gonzalez, A., … (2022). A gene signature... Cancers, 14, 136. 10.3390/CANCERS14010136/S1
- Bueno-Fortes, S., Berral-Gonzalez, A., …(2023). Identification of a gene ... Bioinformatics Advances, 3. 10.1093/BIOADV/VBAD037
- D. R. Cox (1972), Regression Models and Life-Tables, *JRSS*-2, 187–202, 10.1111/j.2517-6161.1972.tb00899.x
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for... Journal of Statistical Software, 33, 1. 10.18637/jss.v033.i01
- Guinney, J., … (2015). The consensus molecular subtypes of colorectal cancer. Nature Medicine, 21, 1350–1356. 10.1038/nm.3967
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased … Technometrics, 12(1), 55–67. 10.1080/00401706.1970.10488634
- Kaplan, E. L., & Meier, P. (1958). Nonparametric… JASA 53(282), 457–481. 10.1080/01621459.1958.10501452
- Mantel N (1966). Evaluation of survival data and two new rank…Cancer Chemotherapy Reports, 50, 163–170
- Quiroga, M.,…(2022). Protein degradation by e3 ubiquitin ligases in cancer stem cells. Cancers, 14, 990. 10.3390/CANCERS14040990
- Rodríguez-Alonso, A., … (2020). Regulation of epithelial–mesenchymal … Cancers, 12, 3093. 10.3390/CANCERS12113093
- Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso, JRSS, 58, I-1,1996, 267–288, 10.1111/j.2517-6161.1996.tb02080.x
- Tibshirani, R. J. (2009). Univariate shrinkage in the cox model for high dimensional data. SAGMB, 8. 10.2202/1544-6115.1438
- Zou H, Hastie T, (2005). Regularization and Variable Selection Via the Elastic Net, JRSS, B67, I-2, 301–320, 10.1111/j.1467-9868.2005.00503.x
- StatQuest. Starmer, J. (2022). StatQuest Youtube channel. Available online at: https://www.youtube.com/c/joshstarmer.
- Lifelines (https://lifelines.readthedocs.io/en/latest/Survival%20analysis%20with%20lifelines.html)

Team

Bioinformatics and Functional Genomics

Universidad Pontificia de Salamanca

CSIC
Consejo Superior de Investigaciones Científicas

IBSAL

CIC
Centro de Investigación del Cáncer

VNiVERSiDAD DSALAMANCA

Thanks for your attention

Modeling Survival and Risk

Regression techniques in survival analysis

Dr. José Manuel Sánchez-Santos

Dpto. Estadística - Universidad de Salamanca

Grupo de Bioinformática y Genómica Funcional (Lab 19)

Centro de Investigación del Cáncer (CIC-Usal)