Biostatistics

Applications in Medicine

Nuno Sepúlveda, 03.11.2025

Syllabus

1. General review

    a. What is Biostatistics?
    b. Population/Sample/Sample size
    c. Type of Data – quantitative and qualitative variables
    d. Common probability distributions
    e. Work example – Malaria in Tanzania

2. Applications in Medicine

    a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve,Rogal-Gladen estimator
    b. Estimation of treatment effects - generalized linear models
    c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model

3. Applications in Genetics, Genomics, and other 'omics data

    a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
    b. Methylation association studies – M versus beta values, estimation of biological age
    c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial

4. Other Topics

    a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
    b. Serological analysis – Gaussian (skew-normal) mixture models
    c. Advanced sample size and power calculations

Exercise:

Covariates: Age, Gender, Infection trigger, Disease Duration

Use logit, probit, cloglog, loglog, cauchit.

Compare models/Use a feature selection strategy

Packages ordinal, glmx, and MASS

PLOS | ONE

RESEARCH ARTICLE

B-Lymphocyte Depletion in Myalgic Encephalopathy/ Chronic Fatigue Syndrome. An Open-Label Phase II Study with Rituximab Maintenance Treatment

Øystein Fluge[1]*, Kristin Risa[1], Sigrid Lunde[1], Kine Alme[1], Ingrid Gurvin Rekeland[1], Dipak Sapkota[1,2], Einar Kleboe Kristoffersen[3,4], Kari Sørland[1], Ove Bruland[1,5], Olav Dahl[1,4], Olav Mella[1,4]*

1 Department of Oncology and Medical Physics, Haukeland University Hospital, Bergen, Norway,
2 Department of Clinical Medicine, University of Bergen, Haukeland University Hospital, Bergen, Norway,
3 Department of Immunology and Transfusion Medicine, Haukeland University Hospital, Bergen, Norway,
4 Department of Clinical Science, University of Bergen, Haukeland University Hospital, Bergen, Norway,
5 Department of Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

CrossMark
click for updates

What will be your final model to understand the effect of treatment better?

Penalised regression

Estimation

$\updownarrow$

Model selection


Accuracy
Bias

Penalised regression

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_i \right)^2 \right\} .$$

subject to a constraint

$$pen \leq \lambda$$

pen = penalty function

λ = tuning parameter

Ridge Regression

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_i \right)^2 \right\} ,$$

subject to $\sum_{j=1}^{p} b_j^2 \leq \lambda_2$

$$\lambda_2 \in \left[0, \sum_{j=1}^{p} (\hat{b}_j^*)^2 \right]$$

OLS estimates

Hoerl AE, Kennard RW (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12(1): 55-67.

Geometrical interpretation (2D)

$$\sum_{j=1}^{2} b_j^2 \leq \lambda_2$$

$$r^2(\cos^2\theta + \sin^2\theta) \leq \lambda_2$$

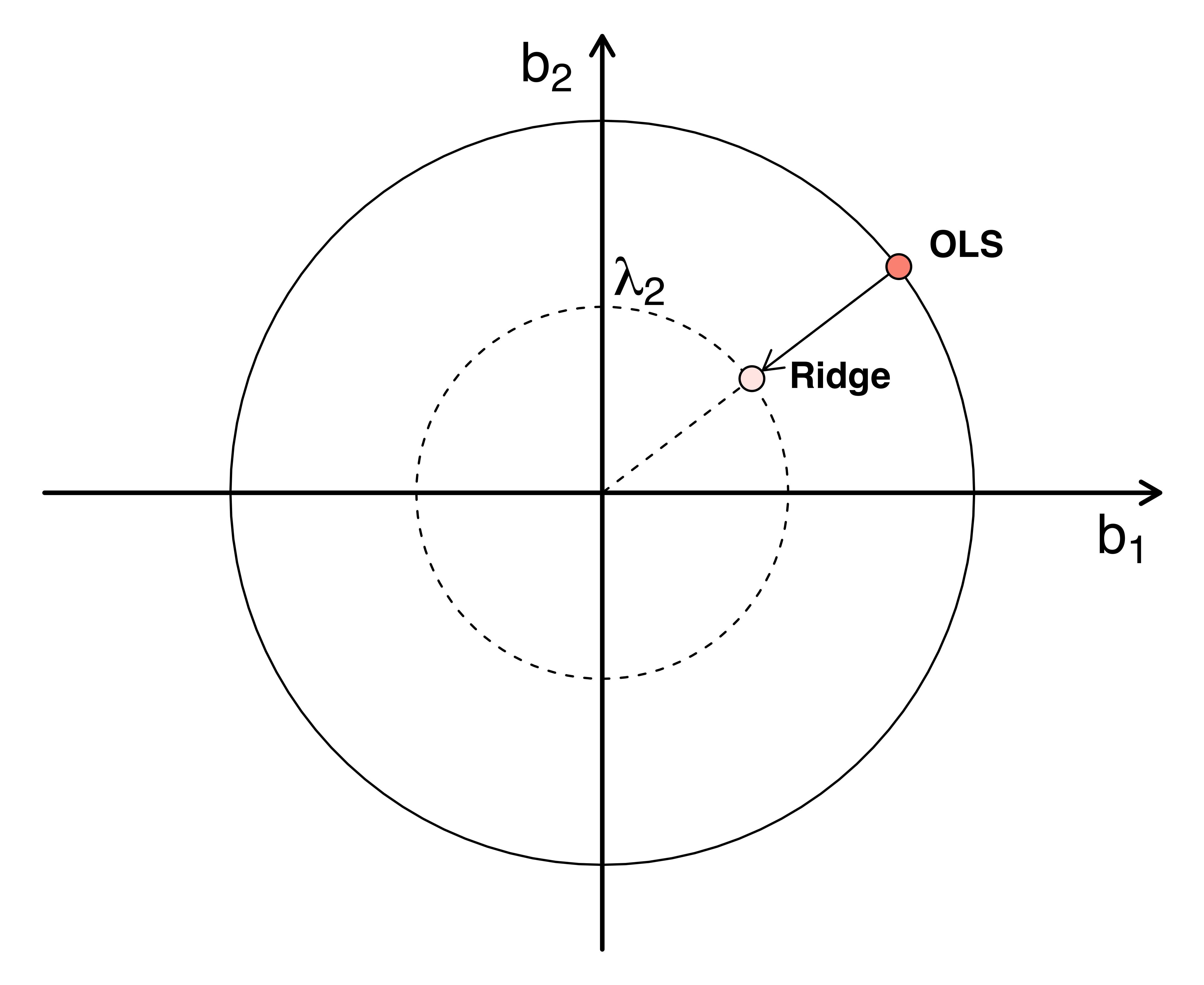
$$r^2 \leq \lambda_2$$

$$b_1 = r\cos\theta$$

$$b_2 = r\sin\theta$$

Ridge estimator is only dependent on the radius and not on the angle

Geometrical interpretation (2D)



Ordinary least squares estimator

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Ridge estimator

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Ridge Regression

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_i \right)^2 \right\} ,$$

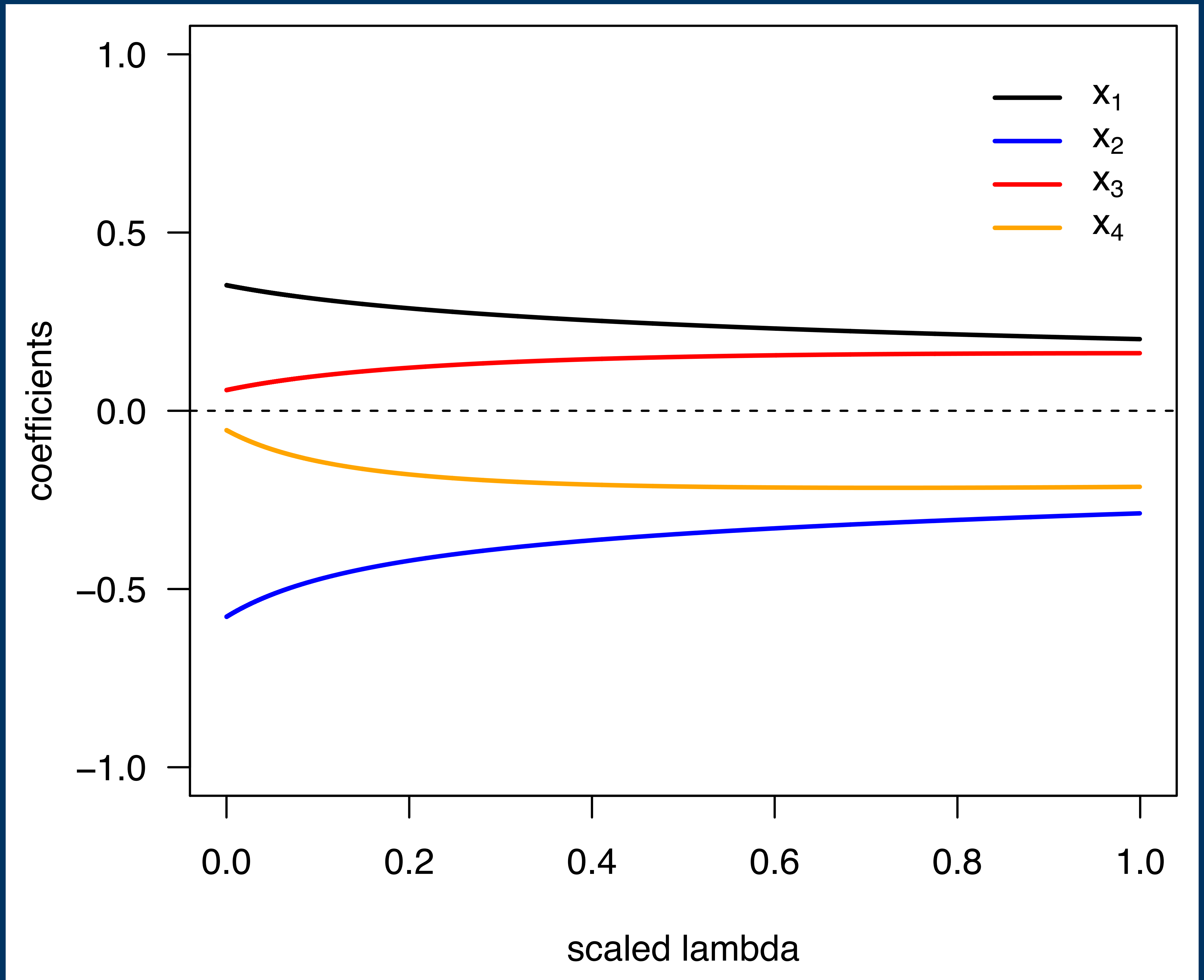
subject to $$\frac{\sum_{j=1}^{p} b_j^2}{\sum_{j=1}^{p} (\hat{b}_j^*)^2} \leq 1 - \lambda^*$$

0% shrinkage

$$\lambda^* \in [0,1]$$

"100%" shrinkage

Ridge trace plot



Ridge regression

Advantages

Remove multicollinearity

Estimator with a closed form

Shrinkage

Disadvantages

Biased estimators

No shrinkage to zero

(No model selection)

LASSO Regression

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_i \right)^2 \right\} ,$$

subject to $\sum_{j=1}^{p} |b_j| \leq \lambda_1$

$$\lambda_1 \in \left[0, \sum_{j=1}^{p} |\hat{b}_j^*| \right]$$

OLS estimates

Geometrical interpretation (2D)

$$\sum_{j=1}^{2} |b_j| \leq \lambda_1$$

$$b_1 = r\cos\theta$$

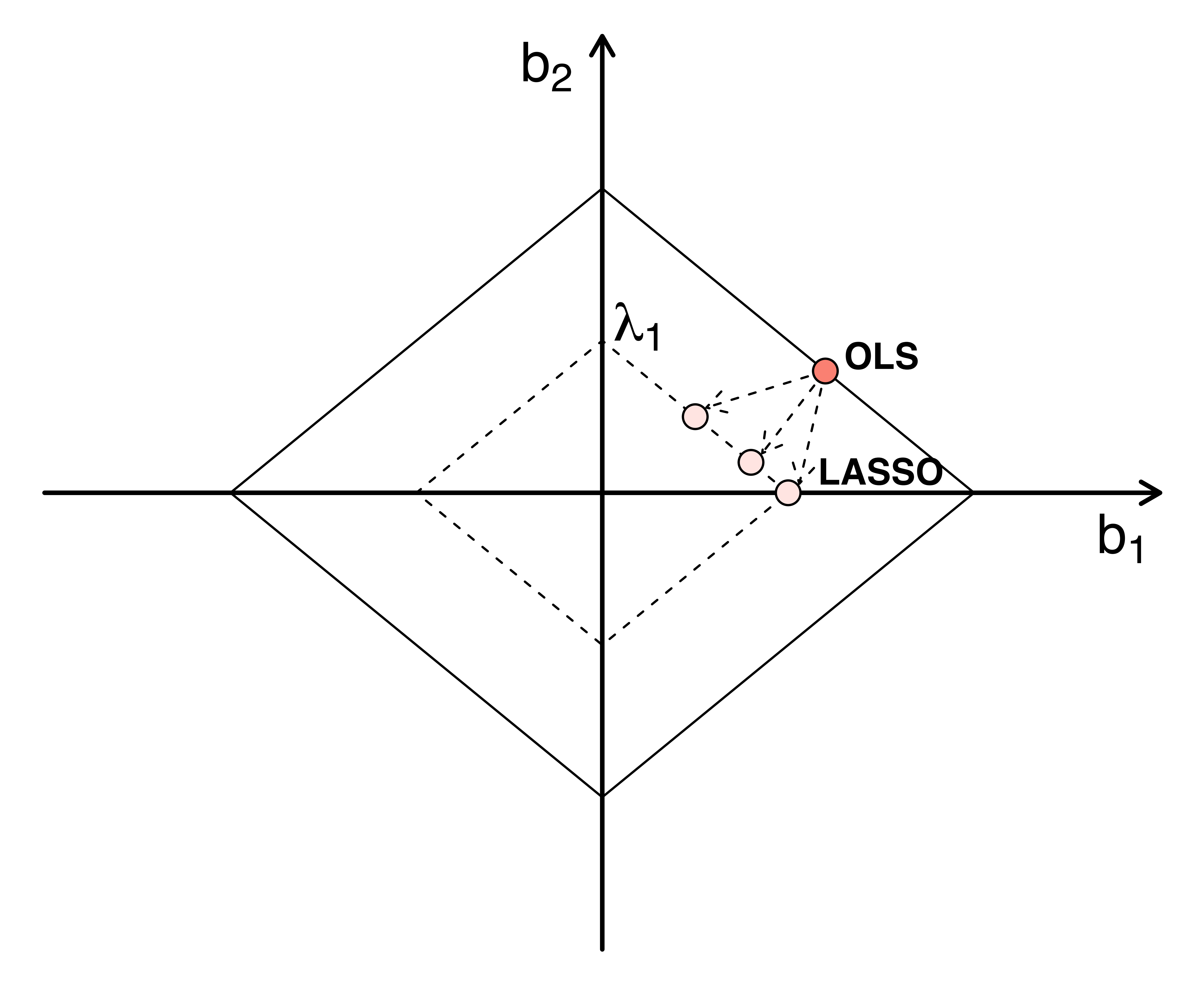
$$b_2 = r\sin\theta$$

$$r(\cos\theta + \sin\theta) \leq \lambda_2$$

$$r^2 \leq \lambda_2$$

LASSO estimator is dependent on both radius and angle

Geometrical interpretation (2D)



LASSO Regression

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_i \right)^2 \right\} ,$$

subject to $$\frac{\sum_{j=1}^{p} |b_j|}{\sum_{j=1}^{p} |b_j^*|} \leq 1 - \lambda^*$$

0% shrinkage (OLS)

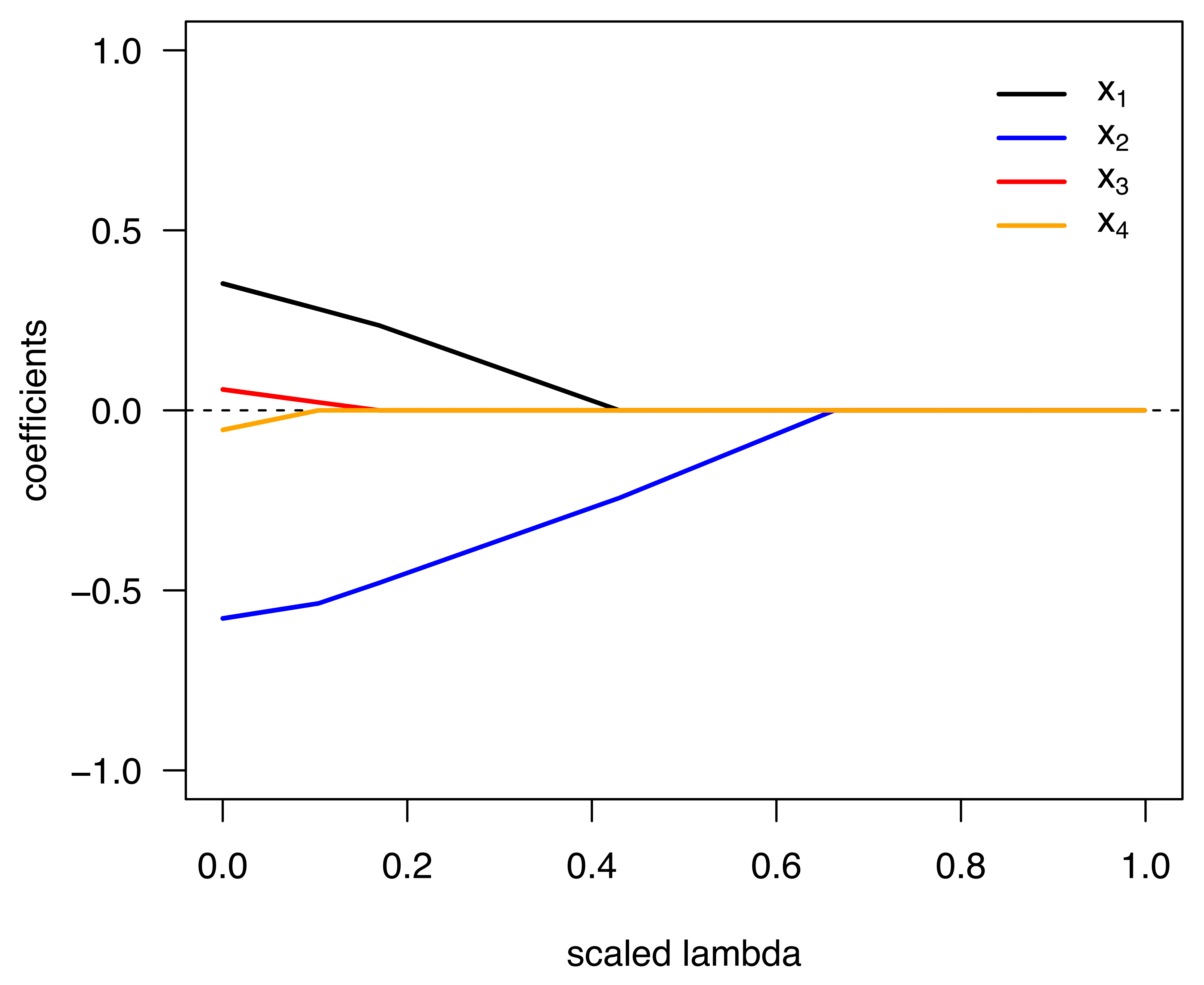
$$\downarrow$$

$$\lambda^* \in [0,1]$$

$$\uparrow$$

100% shrinkage

LASSO trace plot



LASSO regression

Advantages	Disadvantages
Remove multicollinearity	Random choice of highly correlated covariates
Shrinkage to zero	No closed-form expression
(Model selection)	Problems with standard errors

Elastic Net Regression

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_i \right)^2 \right\} \ ,$$

subject to $\alpha |\mathbf{b}|_1 + (1-\alpha)|\mathbf{b}|^2 \leq \lambda$ for some λ and $\alpha \in [0,1]$.

$\alpha = 0 \Rightarrow$ Ridge regression

$\alpha = 1 \Rightarrow$ LASSO regression

Estimation of the tuning parameter(s)

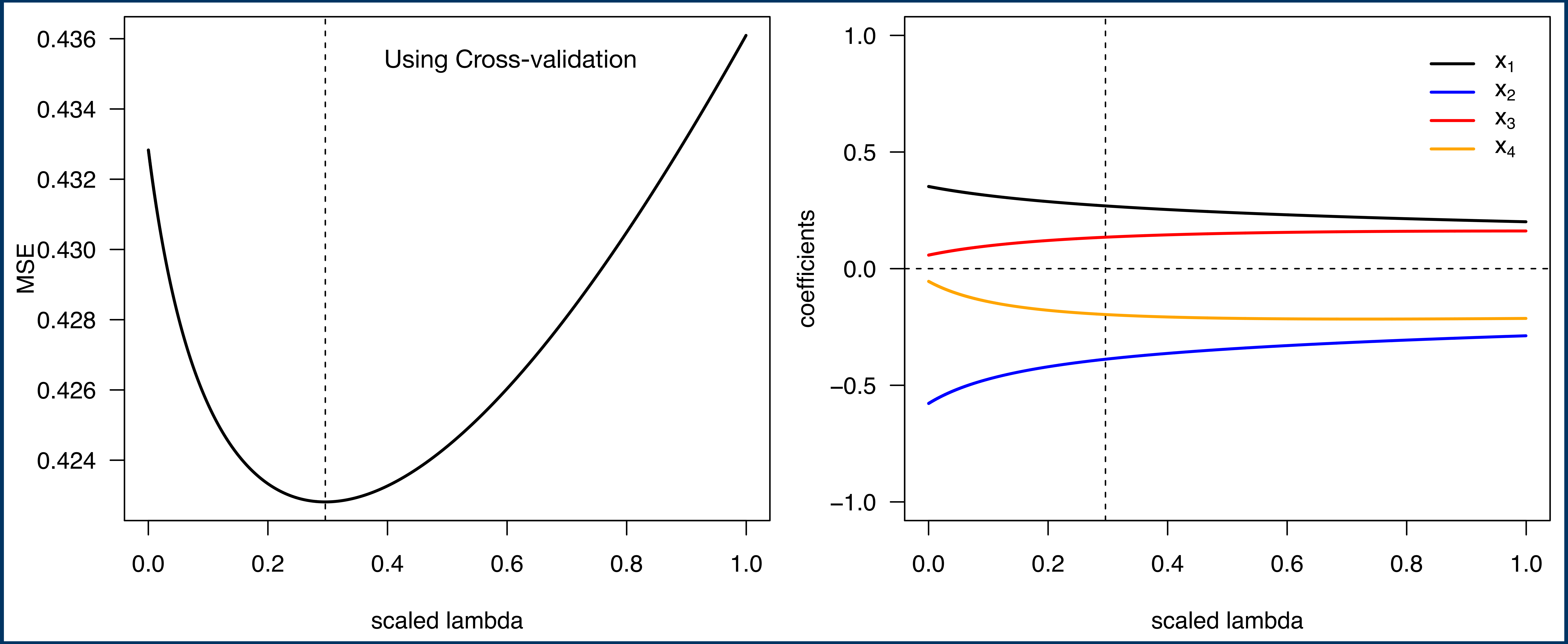
Evaluate a grid of possible values

Highest accuracy

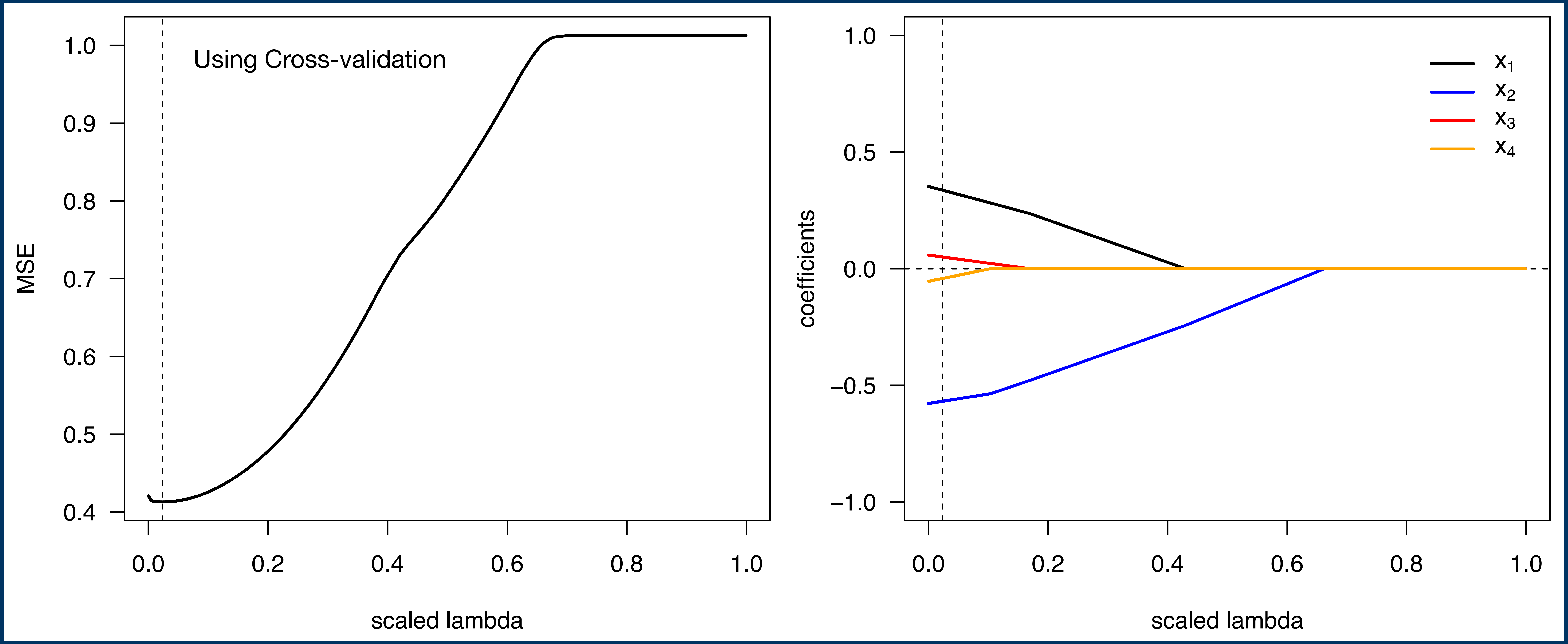
Cross-validation

Lowest mean squared error

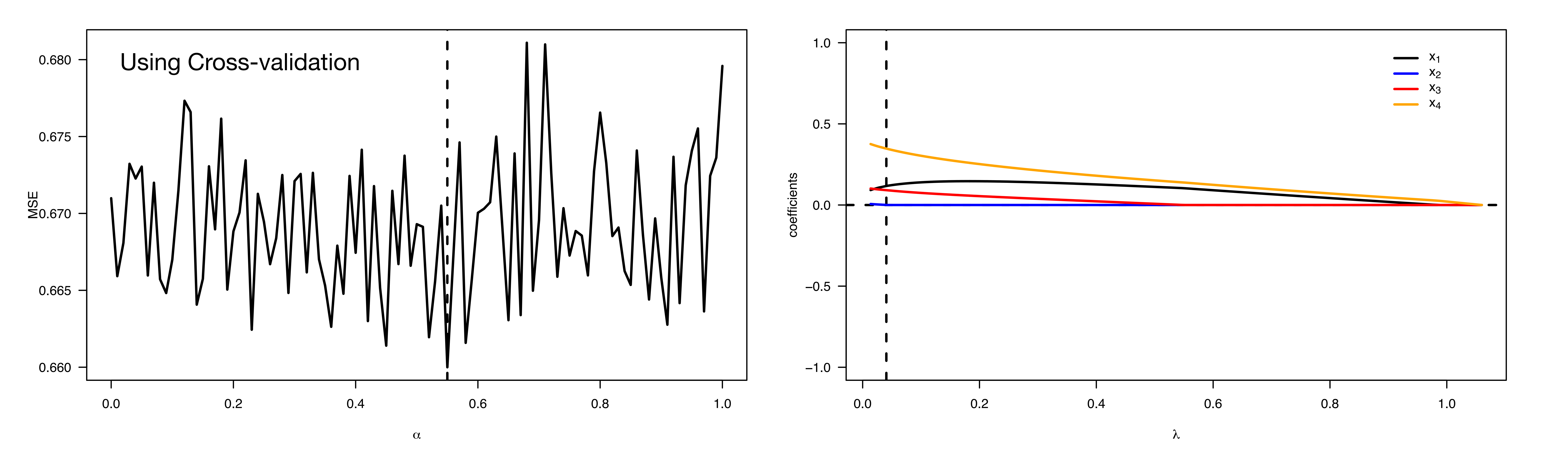
Example: Ridge Regression



Example: LASSO Regression



Example: Elastic Net Regression



Exercise:

Covariates: Age, Gender, Infection trigger, Disease Duration

Package glmnet

Use a binomial model with the probit function

Use LASSO regression

PLOS ONE

RESEARCH ARTICLE

B-Lymphocyte Depletion in Myalgic Encephalopathy/ Chronic Fatigue Syndrome. An Open-Label Phase II Study with Rituximab Maintenance Treatment

Øystein Fluge[1]*, Kristin Risa[1], Sigrid Lunde[1], Kine Alme[1], Ingrid Gurvin Rekeland[1], Dipak Sapkota[1,2], Einar Kleboe Kristoffersen[3,4], Kari Sørland[1], Ove Bruland[1,5], Olav Dahl[1,4], Olav Mella[1,4]*

1 Department of Oncology and Medical Physics, Haukeland University Hospital, Bergen, Norway,
2 Department of Clinical Medicine, University of Bergen, Haukeland University Hospital, Bergen, Norway,
3 Department of Immunology and Transfusion Medicine, Haukeland University Hospital, Bergen, Norway,
4 Department of Clinical Science, University of Bergen, Haukeland University Hospital, Bergen, Norway,
5 Department of Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

What will be the final model to understand the effect of treatment better?

Syllabus

1. General review

    a. What is Biostatistics?
    b. Population/Sample/Sample size
    c. Type of Data – quantitative and qualitative variables
    d. Common probability distributions
    e. Work example – Malaria in Tanzania

2. Applications in Medicine

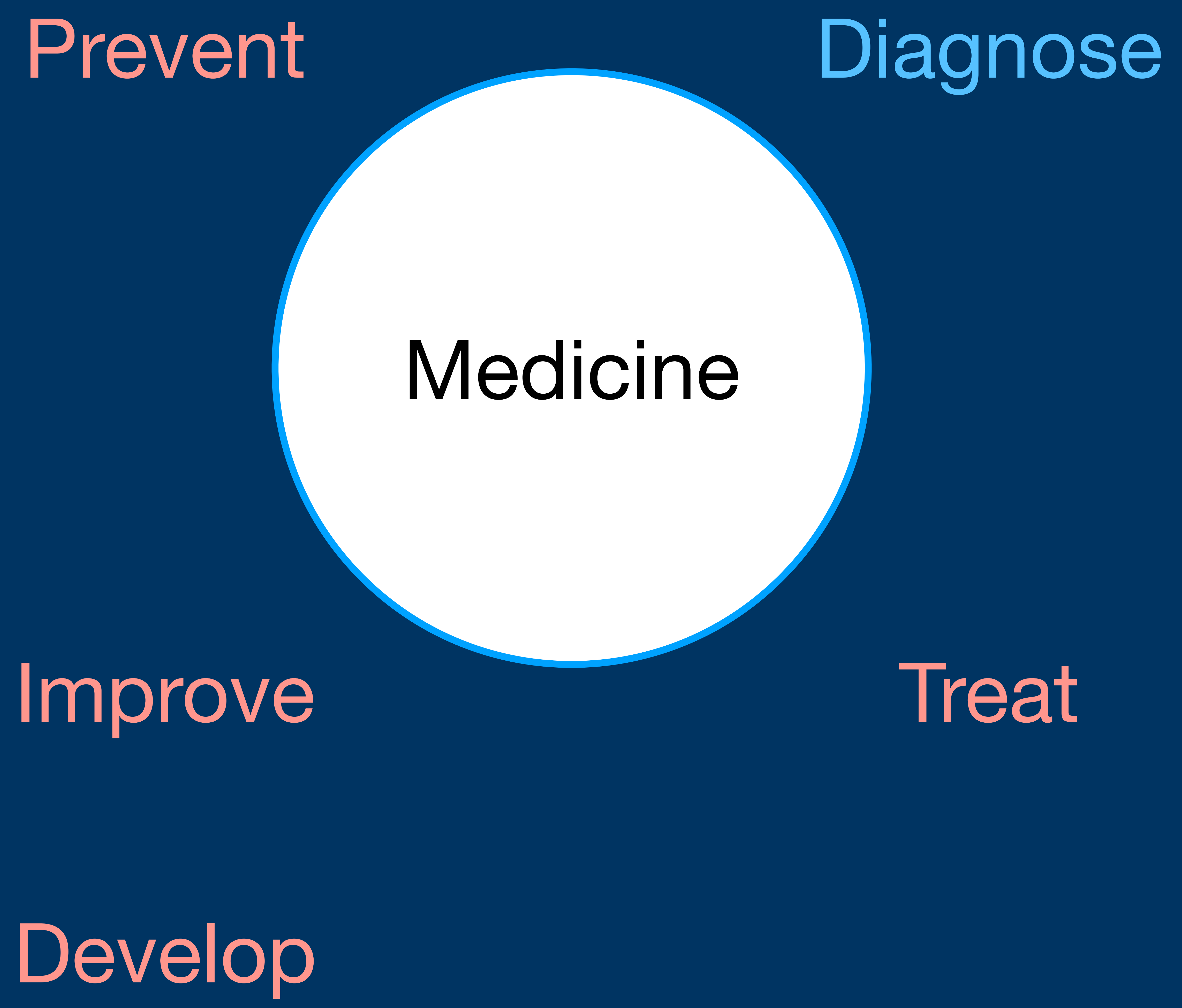
    a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve,Rogal-Gladen estimator
    b. Estimation of treatment effects - generalized linear models
    c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model

3. Applications in Genetics, Genomics, and other 'omics data

    a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
    b. Methylation association studies – M versus beta values, estimation of biological age
    c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial

4. Other Topics

    a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
    b. Serological analysis – Gaussian (skew-normal) mixture models
    c. Advanced sample size and power calculations


Prevent
Diagnose
Medicine
Improve
Treat
Develop

Survival or time-to-event analysis

Endpoint: time to event

Examples of endpoints

time to death in cancer patients (hence, survival analysis)

time to first symptomatic infection after vaccination

time to hospital discharge

time to a positive diagnosis of a chronic disease

time to clearance of infection

What parametric distributions could be used to analyse this random variable?

$T =$ random variable that represents the time when the event of interest occurs

$$T \rightsquigarrow \ ?$$

Survival function

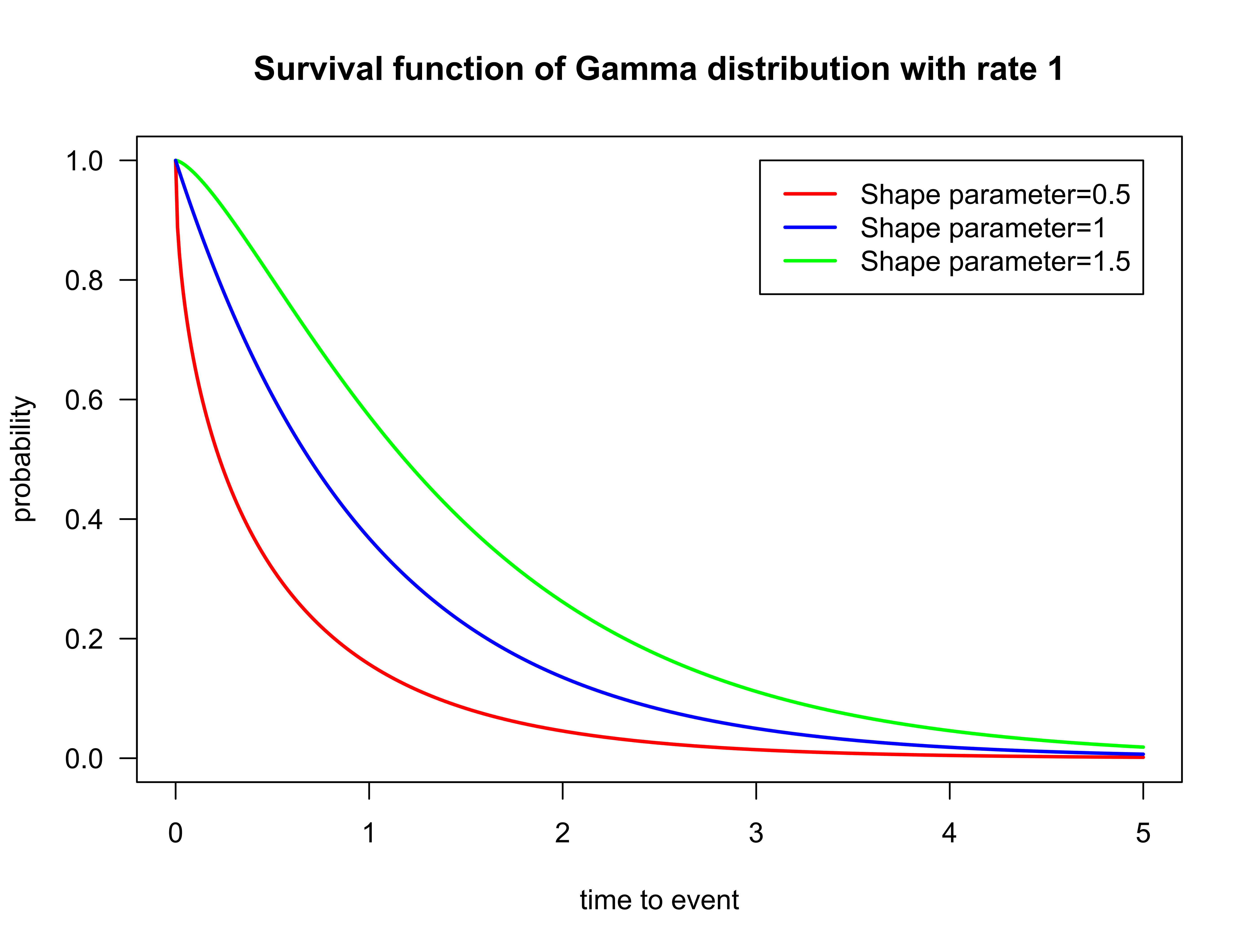
$$S(t) = P(T > t), \; t \geq 0$$

$$S(t) = 1 - F(t), \; t \geq 0$$

S is strictly a decreasing (continuous) function

$S(0) = 1$ and $S(+\infty) = 0$

Example



Survival function of Gamma distribution with rate 1
Shape parameter=0.5
Shape parameter=1
Shape parameter=1.5
1.0
0.8
0.6
0.4
0.2
0.0
probability
0
1
2
3
4
5
time to event

Hazard function (formal definition)

$$h(t) = \lim_{dt \to 0+} \frac{P[t \le T < t + dt \mid T \ge t]}{dt}$$

“Instant” risk of the event occurring at time t

Hazard function (more practical definition)

$$h(t) = \frac{f(t)}{S(t)}$$

Hazard function is simply the ratio between the probability density function and the survival function

Two interesting relationships between probability density function, survival function and hazard function

$$f(t) = \lim_{dt \to 0+} \frac{P[t \leq T < t + dt]}{dt} = -S'(t)$$

$$h(t) = -\frac{S'(t)}{S(t)} \Leftrightarrow S(t) = e^{-\int_0^t h(x)dx}$$

(by the fundamental theorem of calculus)

Exercise 0

Use the practical definition of hazard function and plot the hazard functions of the following distributions:

Exponential distribution with rate parameter =1

Gamma distribution with shape parameter = 0.5 and rate parameter =1

Gamma distribution with shape parameter = 1.5 and rate parameter =1

What is your interpretation of these hazard functions?

Discussion

What is the qualitative aspect of the hazard function for time to death in humans?

Exercise 1: data about recovery from a SARS-CoV-2 infection

16 patients from a Beijing hospital between
January 28 and February 9, 2020

time to end of symptoms

time to negative PCR test

Package MASS

Fit exponential, gamma, lognormal, and weibull distributions to each endpoint

Select the best model to each endpoint and plot the corresponding survival and hazard functions

Draw your conclusions

Chang D, Mo G, Yuan X, et al. Time Kinetics of Viral Clearance and Resolution of Symptoms in Novel Coronavirus Infection. Am J Respir Crit Care Med. 2020;201(9):1150-1152.