Biostatistics

Applications in Medicine

Nuno Sepúlveda, 06.11.2023

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

Accuracy

Parsimony

The art of statistical modelling

Multicollinearity

Interpretability

Generalisation

The art of constructing a model

Select the best link function

Fit models with different link functions and compare them

Select the best subset of covariates (feature selection)

Forward/Backward/Stepwise Regression

Penalised regression (LASSO or Elastic-Net)

Classical model comparison and selection

AIC - Akaike's Information Criterion

$$\text{AIC (M)} = (-2)\text{log-L}(\hat{\theta}\,|\,M, \mathbf{x}) + 2p$$

BIC - Bayesian Information Criterion

$$\text{BIC (M)} = (-2)\text{log-L}(\hat{\theta}\,|\,M, \mathbf{x}) + p\log(n)$$

$\text{log-L}(\hat{\theta}\,|\,M, \mathbf{x})$ is the log-likelihood function evaluated on the parameter estimates

p is the number of parameters of model M

n is the sample size

Choose the model with the lowest values of one of these measures

Forward selection

"Empty" Model

Add covariate

Add covariate

Add covariate

⋮

Stop procedure

Increased accuracy compensates increased model complexity

Increased accuracy does not compensate increased model complexity

Backward elimination

"All covariates" Model

Remove covariate

Remove covariate

Remove covariate

Stop procedure

Decreased model complexity does not have an impact on model accuracy

Decreased model complexity has an impact on model accuracy

Stepwise regression

"Empty" Model

Add covariate 1

Add covariate 2
Remove covariate 1

Add covariate 3
Remove covariates 1, 2

⋮

Stop procedure

Increased accuracy compensates increased model complexity

Increased accuracy does not compensate increased model complexity

Stepwise regression

Advantages	Disadvantages
Remove multicolinearity	Overestimation of the number of predictors
Easy automation	Inflated type I errors
Speed	Unstable to slight changes in the data

Exercise:

Covariates: Age, Gender, Infection trigger, Disease Duration

Use logit, probit, cloglog, loglog, cauchit, Aranda-Ordaz link functions

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

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

Penalised regression

Estimation

↕

Model selection


Bias
Accuracy

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)

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

$$b_1 = r\cos\theta$$

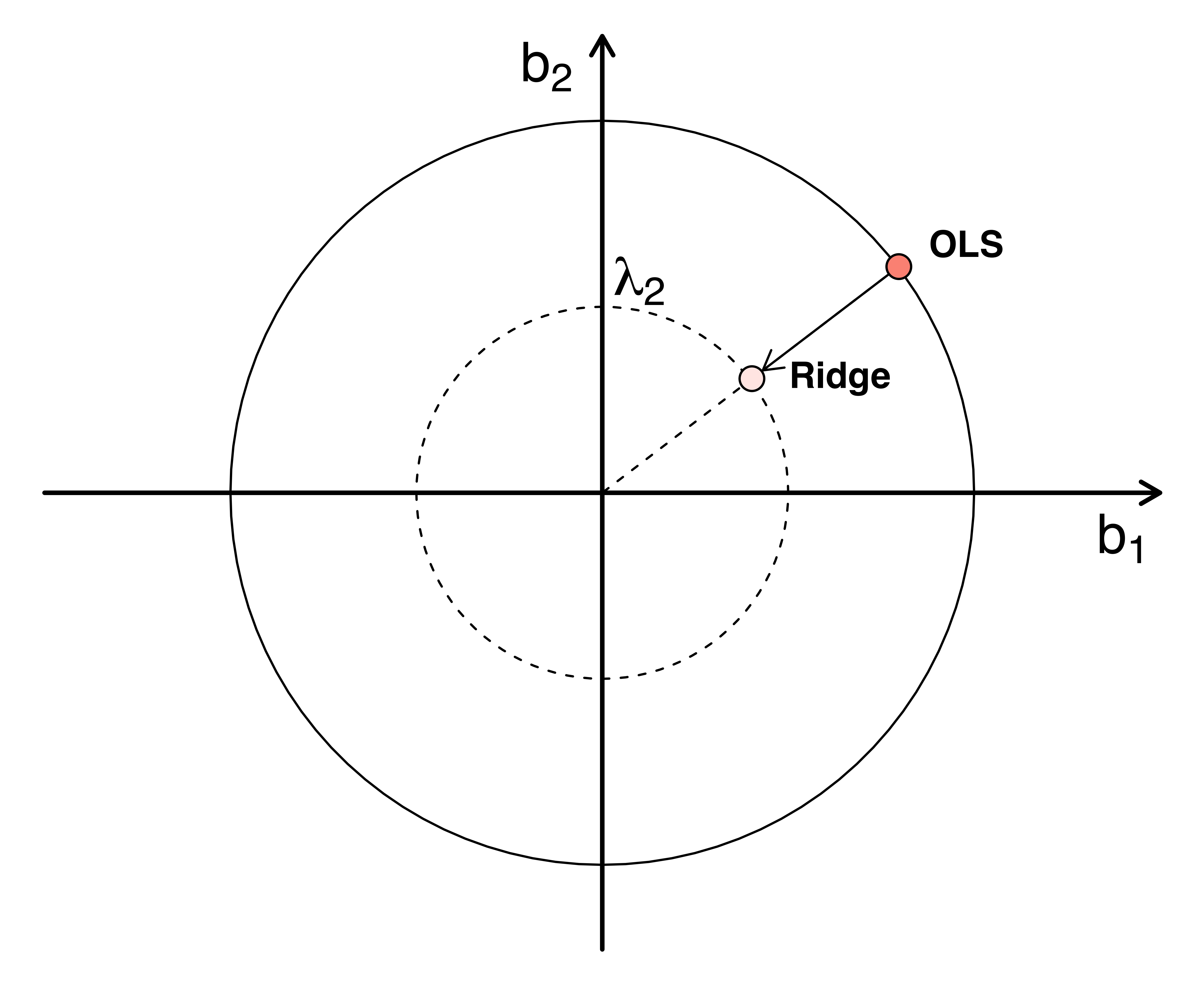
$$b_2 = r\sin\theta$$

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

$$r^2 \leq \lambda_2$$

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

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

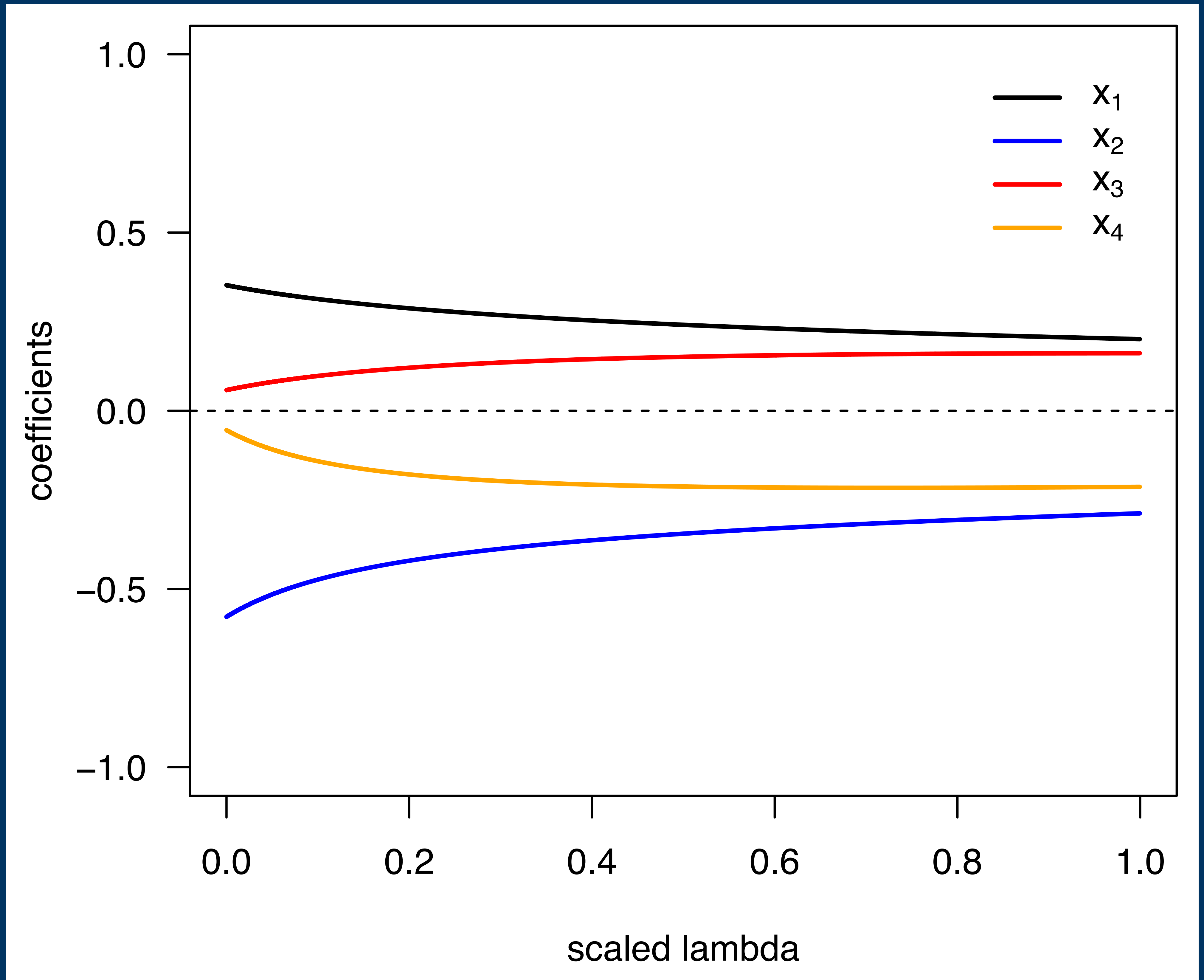
subject to $$\frac{\sum_{j=1}^{p} b_j^2}{\sum_{j=1}^{p} (\hat{b}_j^*)^2} \leq 1 - \lambda^*$$

0% shrinkage

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

“100%” shrinkage

Ridge trace plot



Ridge regression

Advantages	Disadvantages
Remove multicollinearity	Biased estimators
Estimator with a closed form	No shrinkage to zero
Shrinkage	(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)

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

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

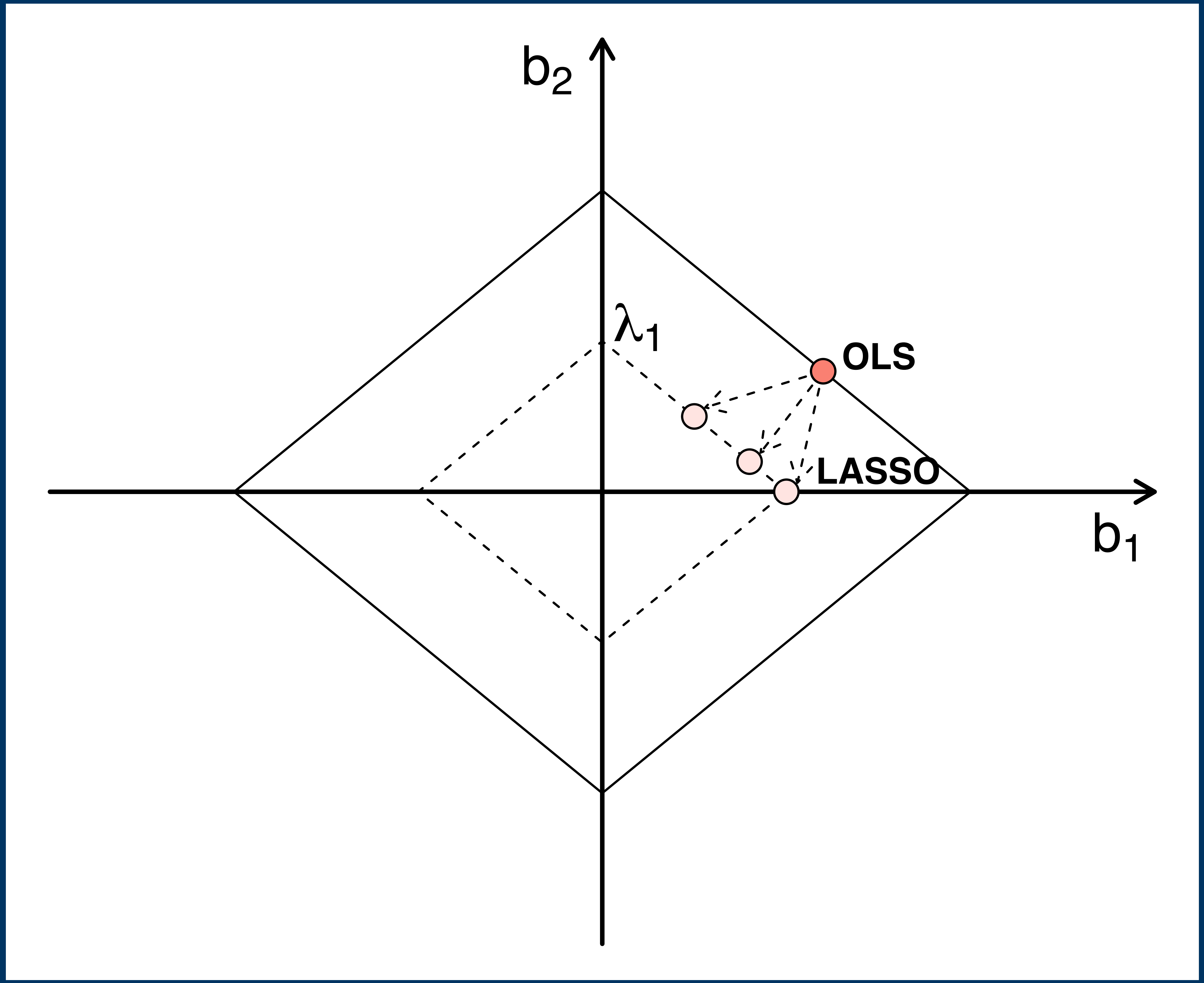
$$r^2 \leq \lambda_2$$

$$b_1 = r\cos\theta$$

$$b_2 = r\sin\theta$$

LASSO estimator is dependent on both radius and angle

Geometrical interpretation (2D)



LASSO Regression

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

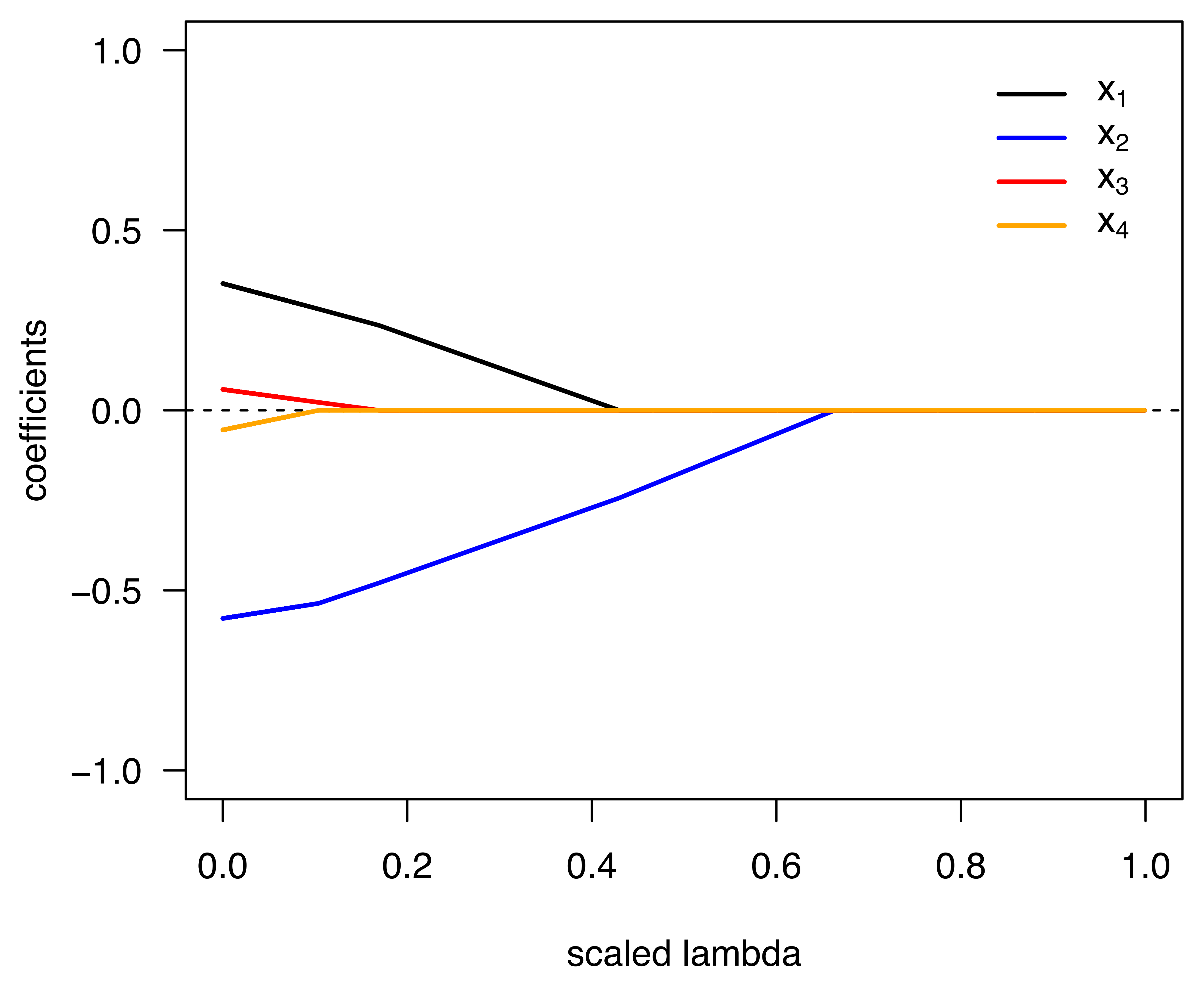
subject to $$\frac{\sum_{j=1}^{p} |b_j|}{\sum_{j=1}^{p} |b_j^*|} \leq 1 - \lambda^*$$

0% shrinkage (OLS)

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

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

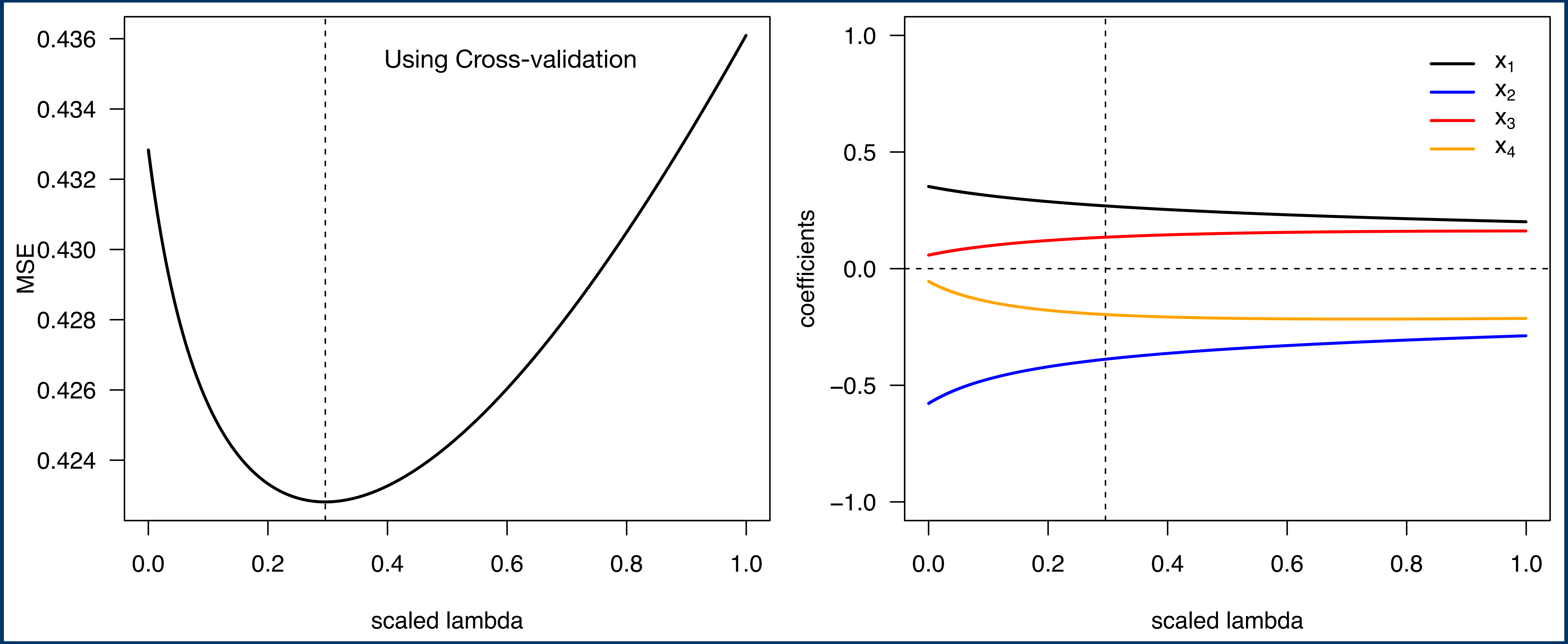
$\alpha = 1 \Rightarrow$ LASSO regression

Estimation of the tuning parameter(s)

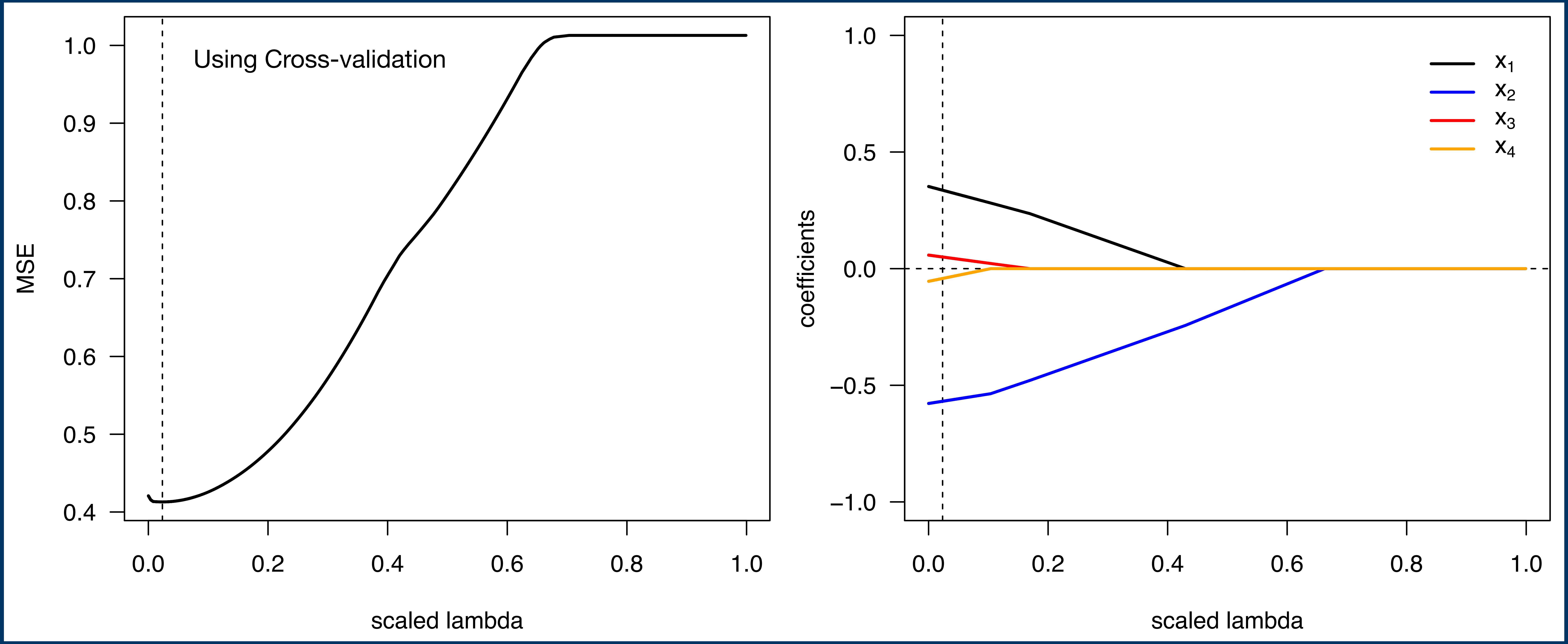
Evaluate a grid of possible values → Highest accuracy ← Cross-validation

Highest accuracy ↕ Lowest mean squared error

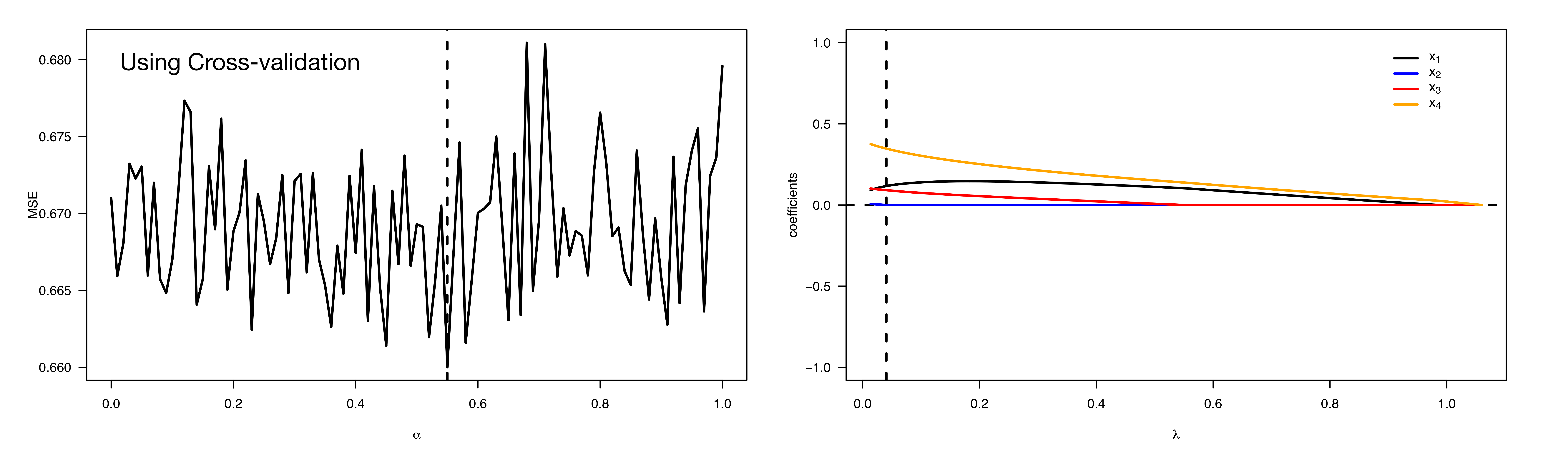
Example: Ridge Regression



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?