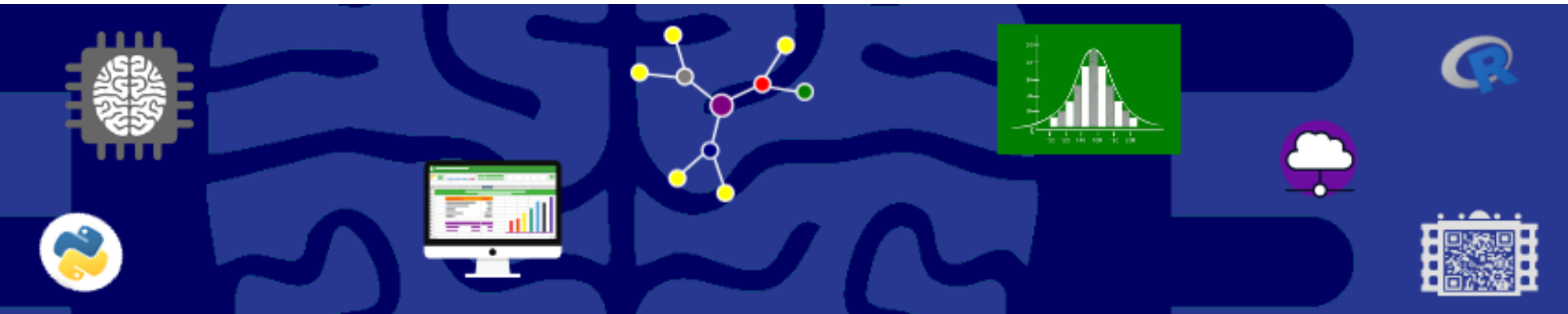
Introduction to Data Science

Winter Semester 2018/19

Oliver Ernst

TU Chemnitz, Fakultät für Mathematik, Professur Numerische Mathematik

Lecture Slides



Contents I

Contents II

Contents III

Contents

Linear Model Selection and Regularization

Chapter overview

- Alternative fitting procedures to least squares (LS) for standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \tag{6.1}$$

  to improve **prediction accurary** and **model interpretability**.
- Prediction accuracy: for approximately linear (true) model, LS has low bias and, if $n \gg p$, also low variance. More variability if $n \gtrsim p$, no unique minimizer if $n < p$.
  Idea: constraining or **shrinking** estimated coefficients reduces variability in these cases at negligible increase in bias, improving prediction accuracy.
- Model interpretability: some predictor variables may be irrelevant for response; LS will not remove these, hence consider other methods for **feature selection** or **variable selection** to exclude irrelevant variables from multiple regression model (by producing zero coefficients for these).

We consider three classes of fitting alternatives to LS:

- **Subset selection**: Find subset of initial p predictor variables which are relevant, fit model using LS for reduced set of variables.
- **Shrinkage**: fit all p variables, shrink coefficients towards zero relative to LS estimate. Shrinkage (also known as **regularization**) reduces variance, some coefficients shrunk to zero, can be viewed as variable selection.
- **Dimension reduction**: project p predictors into subspace of dimension $M < p$, i.e., construct M linearly independent *pseudo-variables* which depend linearly on original p predictor variables. Use these as new predictors for LS fit.
- Same concepts apply to other methods (e.g. classification).

Contents

Idea: Perform separate LS regression for *all possible subsets* of given p predictor variables.

Algorithm 1: Best subset selection.

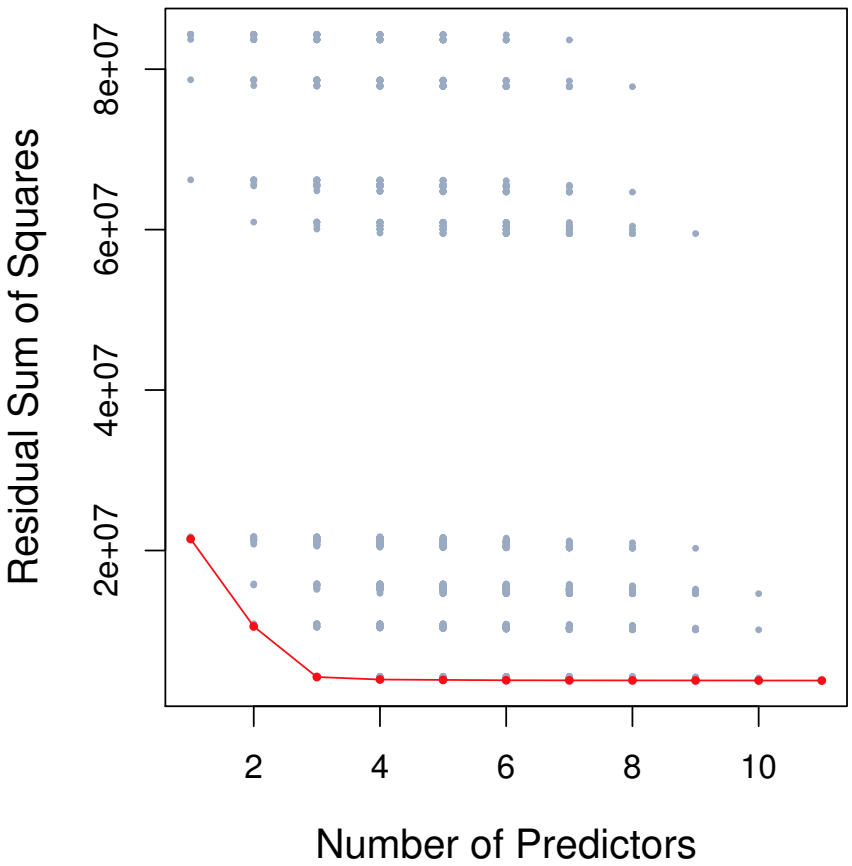
---

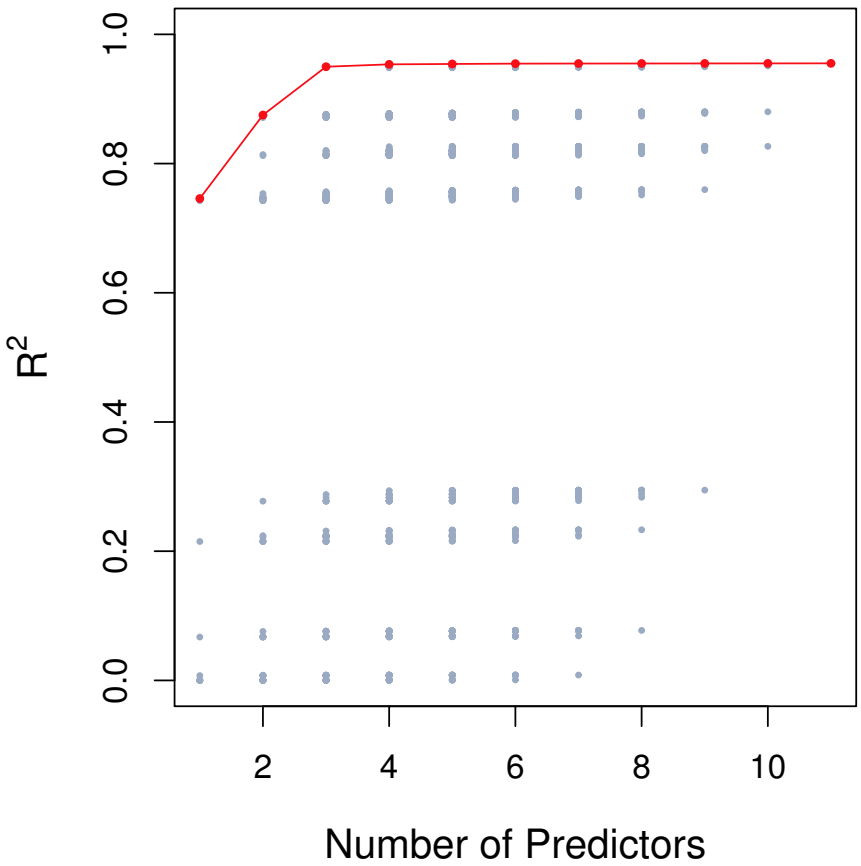
1. Set $\mathscr{M}_0$ to be the **null model**, i.e., containing only constant term β_0.
2. **for** $k = 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models containing exactly k predictors.
   b. Pick best (smallest RSS, i.e., largest R^2) among these, call it $\mathscr{M}_k$.
3. Select single best model among $\mathscr{M}_0, \ldots, \mathscr{M}_p$ using model selection criterion (later).

- Step 2 reduces # model candidates from 2^p to $p + 1$.
- Models in Step 3 display monotone decreasing RSS (increasing R^2) as # variables increases.
- Want low test error rather than low training error.

Linear Model Selection and Regularization

Best subset selection





Best subset selection for `Credit` data set: 10 predictors (three-valued variable `ethnicity` coded using two dummy variables selected separately).

Red line indicates model with smallest RSS (largest R^2).

- Can apply to classification problems using **deviance** in place of RSS ($-2 \cdot$ maximized log-likelihood).
- Best subset selection simple, but # regression fits to compare grows exponentially with p (e.g. 1024 for $p = 10$, over 1 million for $p = 20$).
- Also, statistical problems for large p: the larger the search space, the higher the chance of finding models performing well on training set, but badly for test set.

Idea: Add predictors to model one at a time, at each step adding variable leading to greatest additional improvement.

Algorithm 2: Forward stepwise selection.

---

1. Set $\mathscr{M}_0$ to be the **null model**, i.e., containing only constant term β_0.
2. **for** $k = 0, 1, \ldots, p-1$
   a. Consider all $p-k$ models augmenting $\mathscr{M}_k$ by one additional predictor.
   b. Pick best (smallest RSS, i.e., largest R^2) among these, call it $\mathscr{M}_{k+1}$.
3. Select single best model among $\mathscr{M}_0, \ldots, \mathscr{M}_p$ using model selection criterion (later).

- Rather than 2^p models considered by best subset selection, forward stepwise selection requires only $1 + p(p+1)/2$ LS fits.
  E.g. $p = 20$: 1,048,576 models for best subset selection, 211 models for fowrard stepwise selection.

- Forward stepwise selection not guaranted to find best model out of 2^p possible. E.g. for $p = 3$, best single-variable model could consist of X_1, while best two-variable model consists of X_2, X_3.
- First 4 selected models for best subset selection and forward stepwise selection on `Credit` data set:

# variables	Best subset	Forward stepwise
1	`rating`	`rating`
2	`rating`, `income`	`rating`, `income`
3	`rating`, `income`, `student`	`rating`, `income`, `student`
4	`cards`, `income`	`rating`, `income`
	`student`, `limit`	`student`, `limit`

- Can use forward stepwise selection in high-dimensional case when $n < p$. However, can only construct submodels $\mathcal{M}_0, \ldots, \mathcal{M}_{n-1}$, since LS can uniquely fit at most $n - 1$ variables.

Idea: Begin with full LS model, successively remove least useful predictor.

Algorithm 3: Backward stepwise selection.

1. Set $\mathscr{M}_p$ to be the **full model**, containing all p predictors.
2. **for** $k = p, p-1, \ldots, 1$
   a. Consider all k models containing all but one of the predictors in $\mathscr{M}_k$.
   b. Pick best (smallest RSS, i.e., largest R^2) among these k models, call it $\mathscr{M}_{k-1}$.
3. Select single best model among $\mathscr{M}_0, \ldots, \mathscr{M}_p$ using model selection criterion (later).

- Again only $1 + p(p+1)/2$ model fits.
- No guarantee of finding best model.
- Requires $n > p$.
- **Hybrid** approaches possible, where addition step followed by removal step.

- In best subset selection, forward selection and backward selection, need to choose best among models containing different # variables.
- RSS and R^2 measures will always select model with all p variables.
- Goal: select best model with respect to *test* error.
- Two basic approaches:
  1. Indirectly estimate test error by making an adjustment to training error to account for bias due to overfitting.
  2. Directly estimate test error using either validation set approach or cross-validation approach.

- Training set MSE generally underestimates test MSE (recall $\mathsf{MSE} = \mathsf{RSS}/n$)
- For LS regression: coefficients determined by minimization of RSS.
- Therefore training error decreases as variables added to model; not so for test error.
- For fitted LS model containing d predictors, C_p **estimate** defined by

$$C_p := \frac{1}{n}(\mathsf{RSS} + 2d\,\hat{\sigma}^2), \tag{6.2}$$

where $\hat{\sigma}^2$ is an estimate of $\mathbf{Var}\,\varepsilon$, typically computed using full model.
Adds penalty term $2d\,\hat{\sigma}^2$ to training RSS to compensate for underestimating test error.
Can show: C_p unbiased estimate of test MSE if $\hat{\sigma}^2$ unbiased estimate of σ^2.
Hence C_p small for models with small test MSE.

- **Akaike information criterion** (AIC) defined for models fit by maximum likelihood.
- For standard linear model (6.1) with Gaussian noise maximum likelihood fit coincides with LS fit.
- In this case

$$\mathsf{AIC} = \frac{1}{n\,\hat{\sigma}^2}(\mathsf{RSS} + 2d\,\hat{\sigma}^2)$$

  (have omitted additive constant).
- Hence, for LS models C_p and AIC proportional.

- **Bayesian information criterion** (BIC), derived from Bayes point of view, is given by (up to irrelevant constants)

$$\mathrm{BIC} = \frac{1}{n\,\hat{\sigma}^2}(\mathrm{RSS} + d\,\hat{\sigma}^2 \log n) \tag{6.3}$$

- Also tends to be small for models with small test error.
- Replaces $2d\hat{\sigma}^2$ used by C_p with $d\hat{\sigma}^2 \log n$, hence places heavier penalty on models with many variables, results in selection of smaller models than C_p.

- Recall $R^2 = 1 - \mathsf{TSS}\,/\,\mathsf{RSS}$,
  $\mathsf{TSS} = \sum(y_i - \bar{y})^2$ total sum of squares for response.
- R^2 increases as variables added to LS model.
- For LS model with d variables, **adjusted** R^2 statistic given by

$$\text{Adjusted } R^2 := 1 - \frac{\mathsf{RSS}\,/(n-d-1)}{\mathsf{TSS}\,/(n-1)} = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}} \cdot \frac{n-1}{n-d-1}. \qquad (6.4)$$

- Unlike C_p, AIC and BIC, where small value indicates model with low test error, here a *large* value of the adjusted R^2 statistic indicates a model with a small test error.
- Maximizing adjusted R^2 equivalent to minimizing $\mathsf{RSS}\,/(n-d-1)$.
- Intuition: once all relevant variables have been included, adding additional noise variables will only lead to small decrease in RSS.
- Compared to R^2, adjusted R^2 pays a price for adding irrelevant variables.

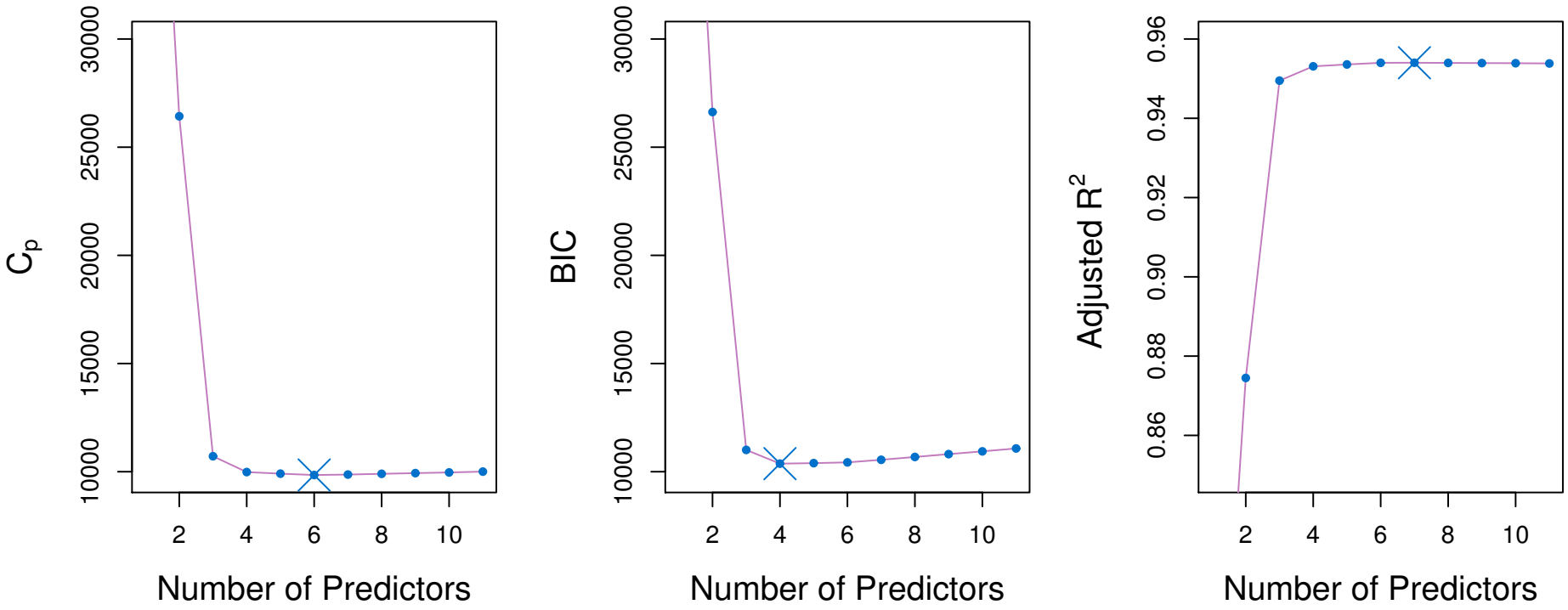
Linear Model Selection and Regularization

C_p, AIC, BIC, adjusted R^2

- Rigorous justifications of C_p, AIC, BIC rely on asymptotic arguments (large n limit).
- Adjusted R^2 popular, intuitive, but not as well motivated statistically.
- All measures simple to use and compute.
- Modified formulas for more general models.

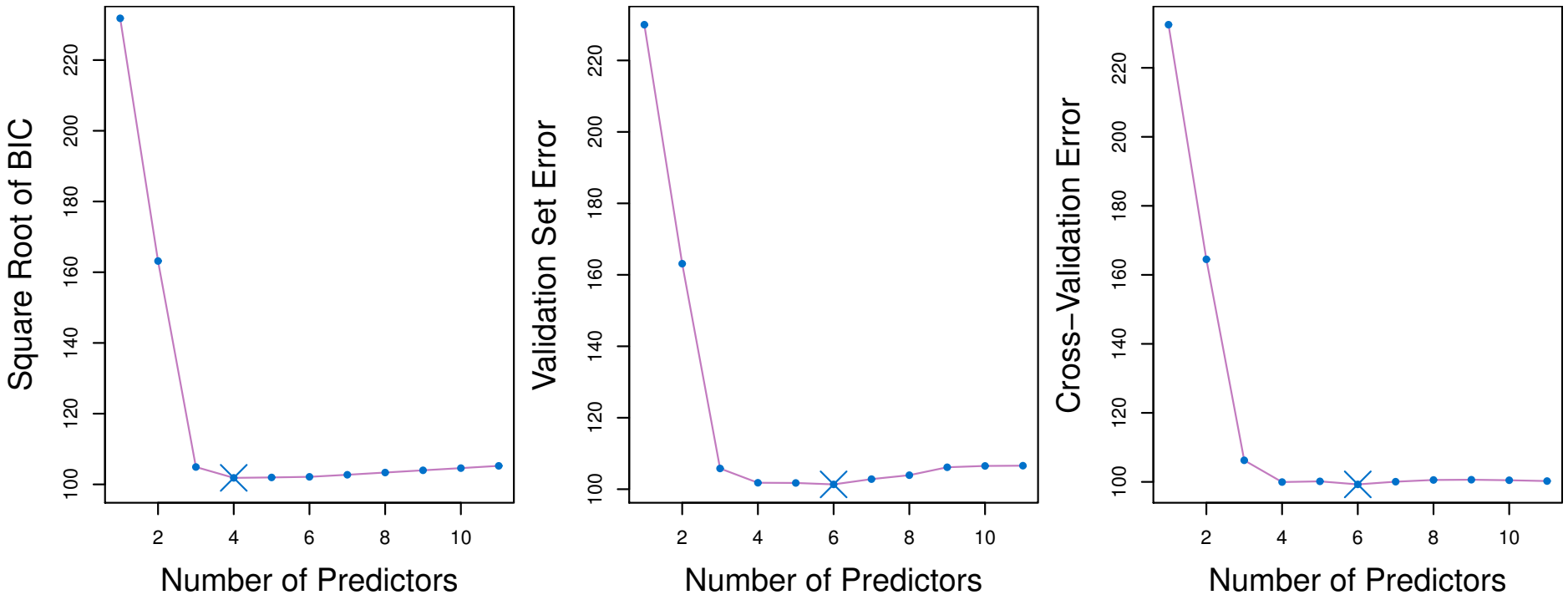
Linear Model Selection and Regularization

C_p, AIC, BIC, adjusted R^2



C_p, **BIC** and adjusted R^2 for best models of each size for `Credit` data set (red curve in previous plot).

- Can apply validation and cross-validation to each model and select that with lowest estimate.
- Advantage over C_p, AIC, BIC, adjusted R^2: direct estimate of test error, fewer assumptions about underlying model.
- More widely useable, e.g., when noise variance estimates difficult to obtain.
- CV initially less popular than C_p, AIC, BIC, adjusted R^2 due to computational cost; this is less and less an issue.
- Apply to `Credit` data set: display BIC, validation set errors, cross-validation errors as function of $d = \#$ variables in model.
  Validation: randomly choose 3/4 of observations as training set, remainder as validation set.
  Cross-validation using $k = 10$ folds.



`Credit` data: 3 model error estimates for best model containing 1 to 11 predictors. Both validation set and CV result in 6-variable models. All approaches agree: not much difference in test error for 4, 5, 6-variable models.

Left: $\sqrt{\text{BIC}}$; center: validation set errors; right: CV errors.

- Observation: all 3 error estimates quite flat from 4 variables onward.
- Error estimate-minimizing model likely to change for different partitions of observations or different choice of CV folds.
- **One-standard-error rule**: calculate standard error of estimated test MSE for each model size, then select smallest model for which estimated test error is within one standard error of lowest point on curve.
  Rationale: if several models appear equally good, may as well choose simplest.
  Here: rule leads to 3-variable model.

Contents

- **Inverse problems**: branch of applied mathematics for solving problems where solution extremely sensitive to data and/or solution not unique (e.g.: X-ray tomography, image deblurring).
- Prevalent strategy: instead of original problem, solve *nearby* problem with better stability properties: **regularization**.
- In LS methods: modify objective function by minimizing different norm or adding **penalty term**, thus imposing "a priori information" on the coefficients.
- In statistics, particularly in LS regression, regularization is known as **shrinkage**, as certain coefficients are "shrunk" in magnitude relative to their values under LS estimation.
- Here we introduce two popular shrinkage techniques: **ridge regression** and the **LASSO**.

Least-squares fitting determines coefficients $\beta_0, \ldots, \beta_p$ by minimizing

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right)^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

In **ridge regression**, one minimizes instead the objective function

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \|\tilde{\boldsymbol{\beta}}\|_2^2, \tag{6.5}$$

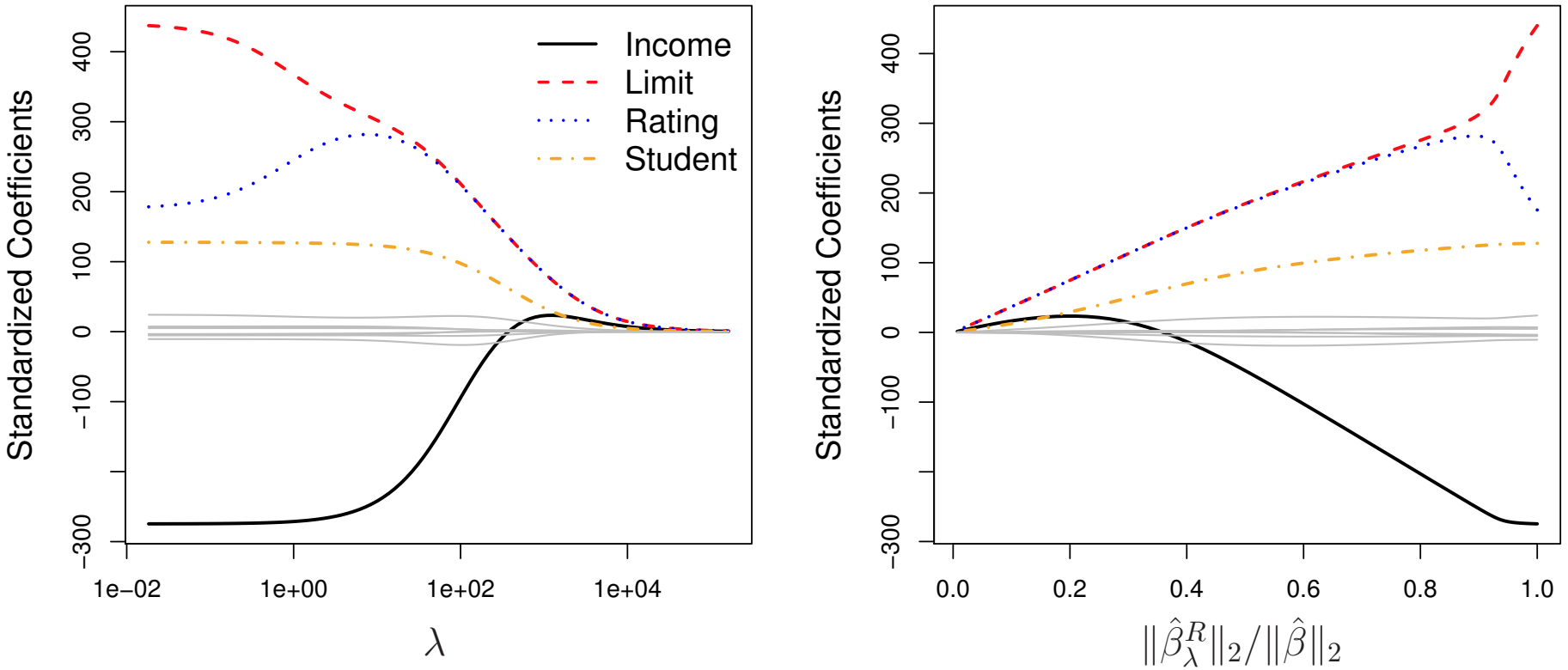
where λ is a tuning parameter to be suitably chosen and $\tilde{\boldsymbol{\beta}} := (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$. From now on $\boldsymbol{\beta} \in \mathbb{R}^p$ and tilde omitted.

In the inverse problems community, this general approach is known as **Tikhonov regularization** and λ is called the **regularization parameter**.

- Tuning λ constitutes tradeoff between two objectives: minimizing RSS (good fit to data) and minimizing **shrinkage penalty** $\lambda\|\boldsymbol{\beta}\|_2^2$, which shrinks $\beta_1, \ldots, \beta_p$ to zero.
- $\lambda = 0$: recover standard LS estimate.
- $\lambda \to \infty$: $\boldsymbol{\beta} \to \mathbf{0}$.
- Different estimate for each value of λ, choice critical.
- Intercept omitted from shrinkage: this is just the mean value of response when all predictor variables are zero.
  Under assumption that all columns of data matrix $\boldsymbol{X}$ have been **centered** to have mean zero, then

$$\hat{\beta}_0 = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

- In the following, for the standard linear model, we tacitly assume $\boldsymbol{X}$ to be centered, the coefficient β_0 to be set to its optimal value $\overline{y}$ and the coefficient vector to be estimated to consist of the components $\beta_1, \ldots, \beta_p$.



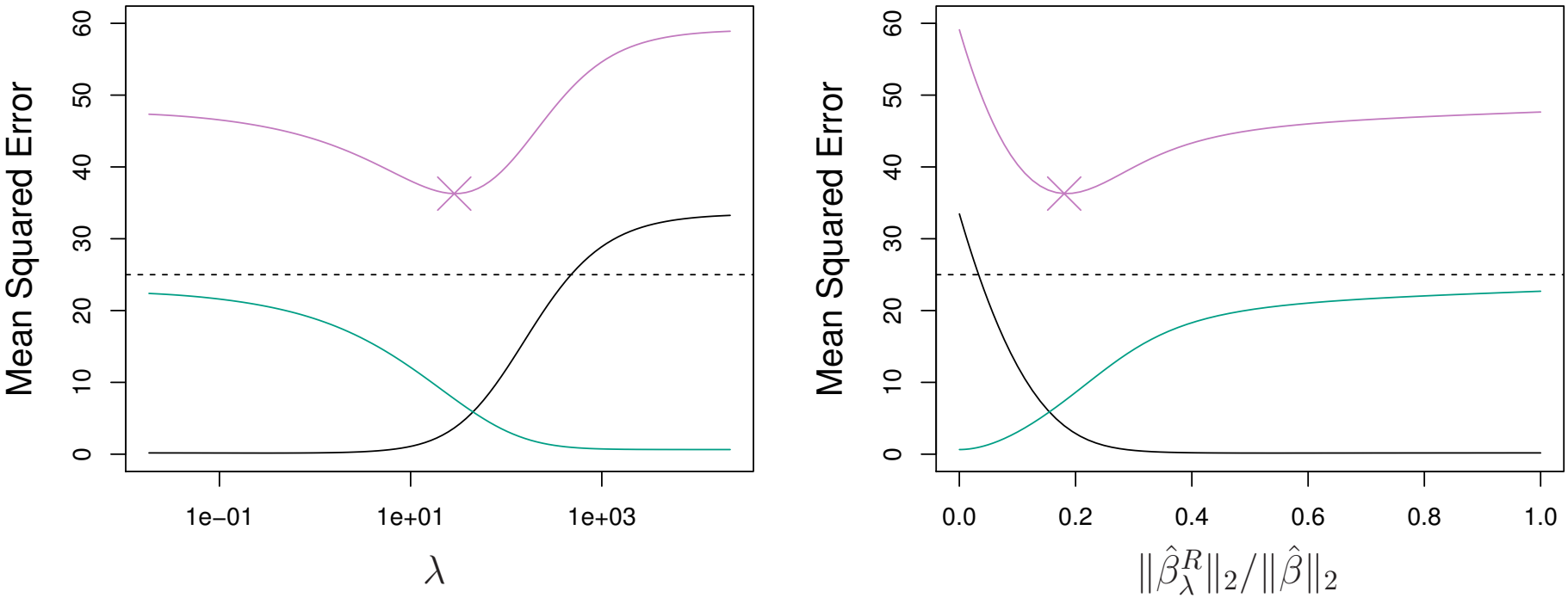
Ridge regression applied to `Credit` data set: values of coefficients of the 10 predictor variables against λ. Lines for largest coefficients `income`, `limit`, `rating` and `student` displayed in distinct colors. Right: x-axis is $\|\hat{\boldsymbol{\beta}}_\lambda^R\|_2/\|\hat{\boldsymbol{\beta}}\|_2$ in place of λ.
Predictor variables standardized before carrying out ridge regression.

- For LS estimation of standard linear model, rescaling a predictor variable $X_j \leftarrow cX_j$ simply results in reciprocal rescaling of estimate as $\hat{\beta}_j \leftarrow \hat{\beta}_j/c$. Consequence: $\hat{\beta}_j X_j$, hence data fit, remains the same. This property is called **scale equivariance**.
- This is no longer the case for ridge regression: value of $\hat{\beta}^R_{j,\lambda} X_j$ depends on λ as well as the scaling of X_j (possibly even the scaling of other predictors).
- Therefore, best to **standardize** predictor variables by transformation

$$x_{i,j} \leftarrow \tilde{x}_{i,j} := \frac{x_{i,j}}{s_j}, \qquad s_j := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2}. \tag{6.6}$$

  Denominator s_j estimates variance of j-th predictor.
- Standardized predictor observations have unit variance estimate.

Bias-variance tradeoff: as λ increases, model flexibility decreased, reducing variance, increasing bias.



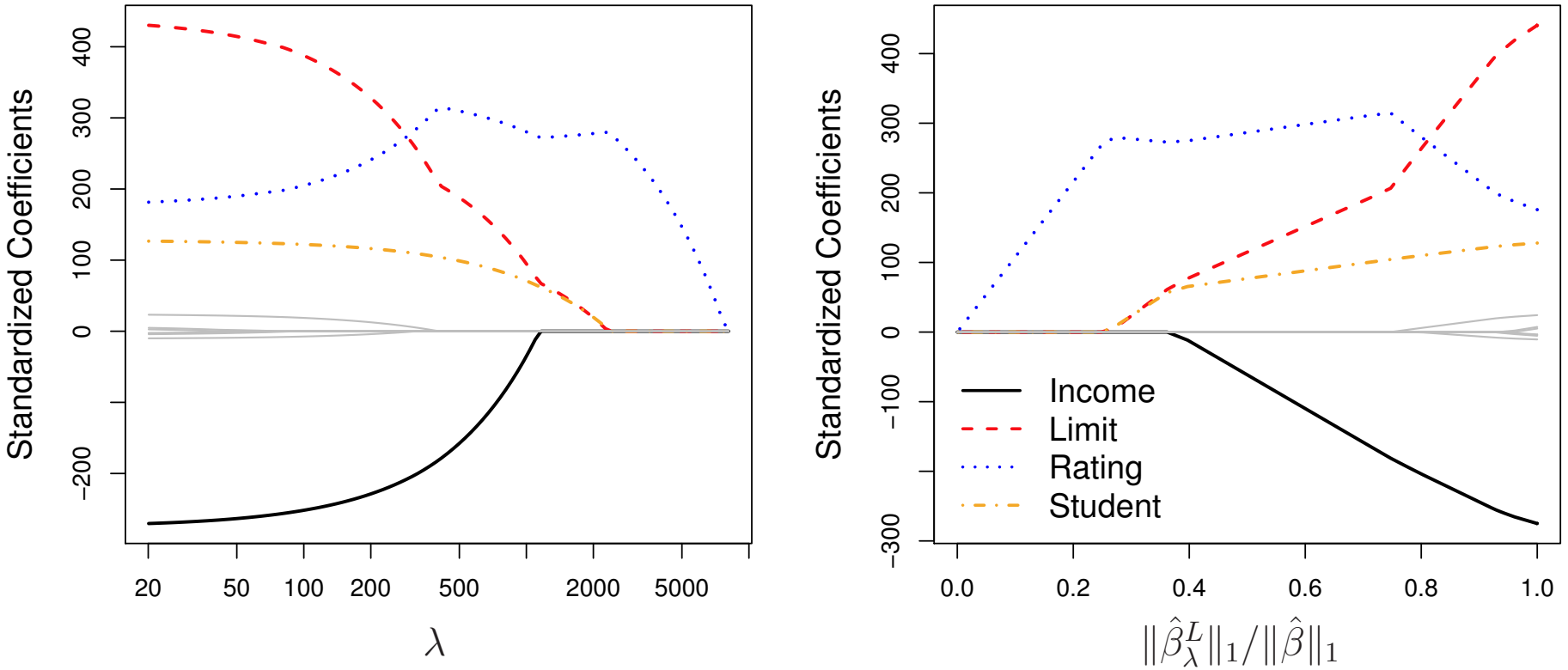
Simulated data, $p = 45$ predictors, $n = 50$ observations. Test MSE (purple), squared bias (black) and variance (green) of ridge regression predictions. Cross: minimal MSE. Dashed line: minimal possible MSE.

- In general: for almost linear (true) model, LS estimate has low bias, but possibly high variance, particularly when p large relative to n.
- For $p > n$ LS fit not unique, but ridge regression still works, trading off slight bias for much reduced variance.
- Computational advantage over best subset selection: ridge regression for many values of λ can be computed at cost of essentially one LS fit, compared to comparing 2^p models.

- Disadvantage of ridge regression: will generally include all p predictors in the model, in contrast with subset selection methods.
- OK for prediction, challenging for interpretation.
- Example: `Credit` data set; most important variables are `income`, `limit`, `rating` and `student`. Model including just these desirable, ridge regression will generally include all 10 predictors.
- **LASSO** (least absolute shrinkage and selection operator): choose coefficients β_j to minimize

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{i,j}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=\mathrm{RSS}+\lambda\|\boldsymbol{\beta}\|_1. \tag{6.7}$$

- ℓ^2-penalty in ridge regression replaced by ℓ^1-penalty, $\|\boldsymbol{\beta}\|_1=|\beta_1|+\cdots+|\beta_p|$.
- ℓ^1-penalty: for λ sufficiently large, results in some estimates $\hat{\beta}^L_{j,\lambda}$ being exactly zero, effecting an implicit **variable selection**, yielding in **sparse models**, which are easier to interpret.



LASSO applied to `Credit` data set. Note difference to ridge regression for intermediate values of λ: as λ increases, coefficients are successively set to zero, thereby removed from model.

Can show: ridge regression and LASSO estimates solve **constrained minimization problems**

$$\hat{\boldsymbol{\beta}}_\lambda^L = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \quad \text{subject to } \|\boldsymbol{\beta}\|_1 \leq s \qquad (6.8)$$

and

$$\hat{\boldsymbol{\beta}}_\lambda^R = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \quad \text{subject to } \|\boldsymbol{\beta}\|_2^2 \leq s, \qquad (6.9)$$

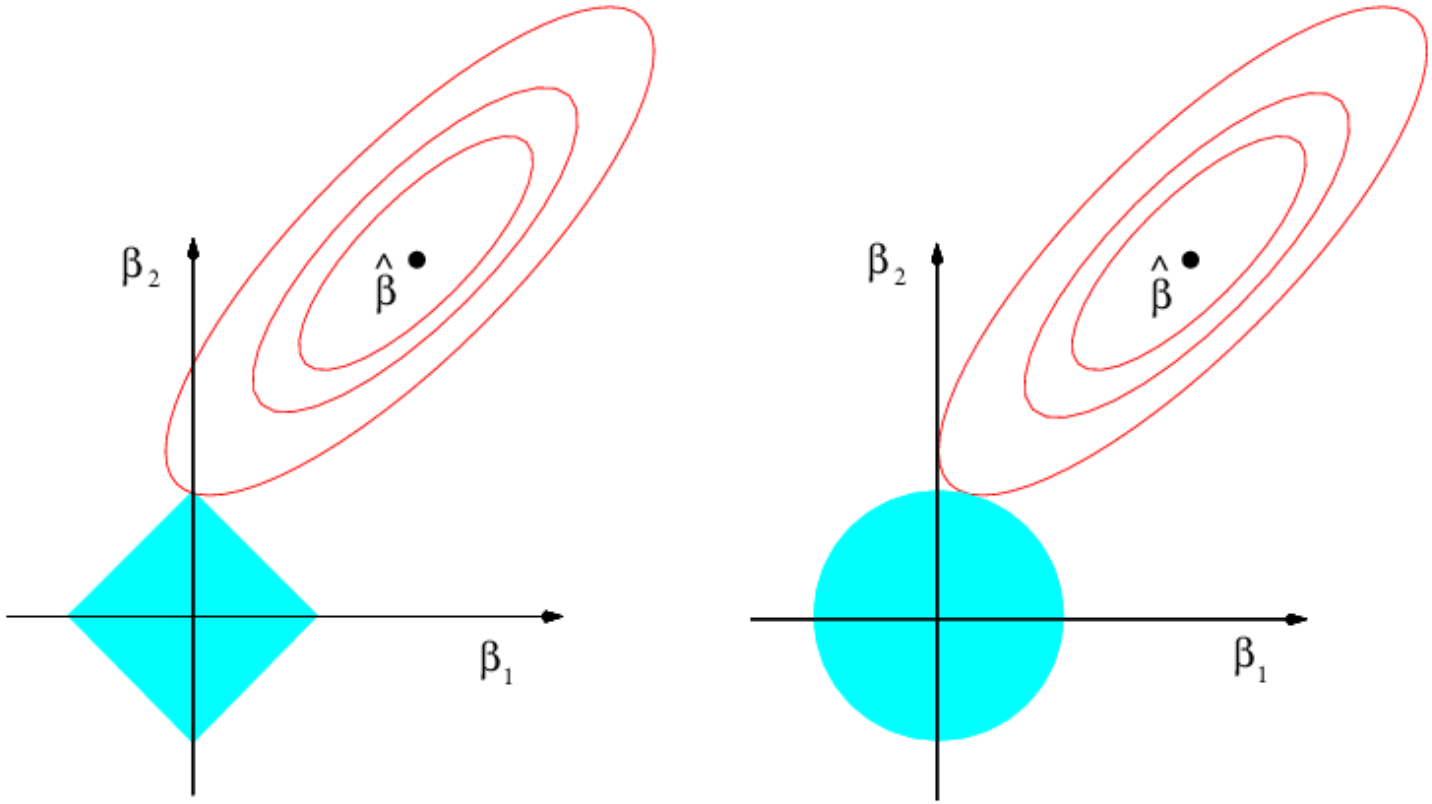
respectively.
In other words: for each value of λ, there is a corresponding value of s, such that both problems give the same estimates.

- Consider constrained minimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \text{ subject to } \sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}} \leq s \quad (6.10)$$

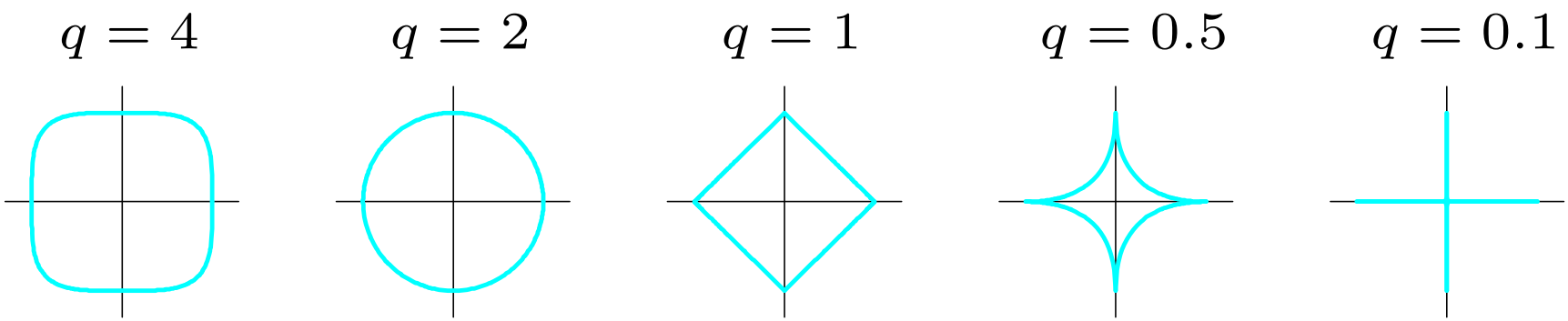
- Minimizes RSS subject to constraint that no more than s coefficients are nonzero.
- This is equivalent to best subset selection.
- Computationally infeasible for large p, since it involves considering all $\binom{p}{s}$ models containing s predictors.
- Hence ridge regression / LASSO computationally feasible alternatives to best subset selection replacing intractable form of budget in (6.10).

- Formulations (6.8) and (6.9) key to understanding variable selection property of LASSO:



Red: RSS contours, blue: constraints $|\beta_1| + |\beta_2| \leq s$ (left) and $\beta_1^2 + \beta_2^2 = s$ (right).

- Unit spheres of $\sum_{j=1}^{p} |\beta_j|^q$ for $q < 2$ progressively sharper (no longer a norm for $q < 1$).

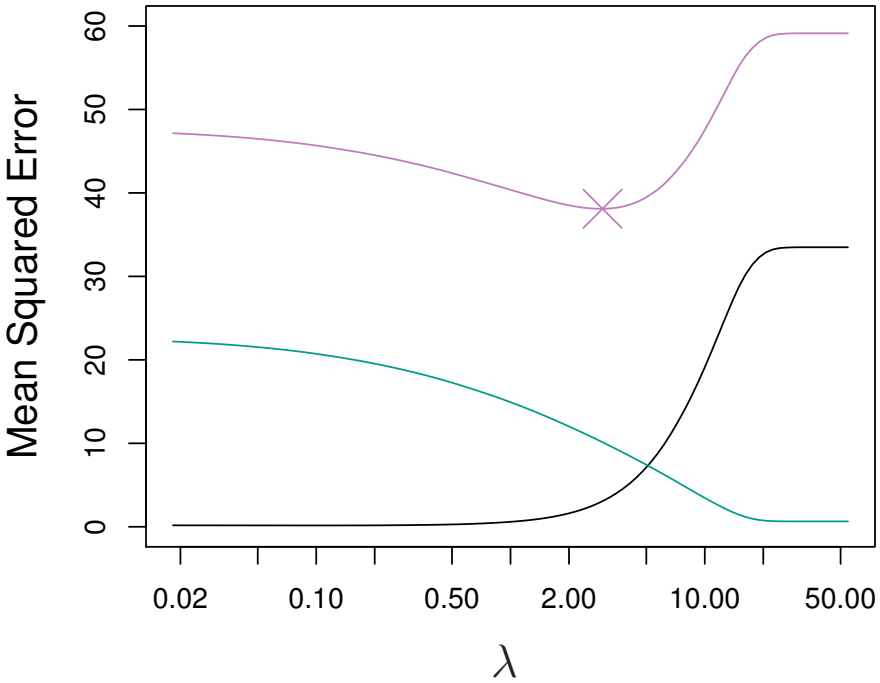


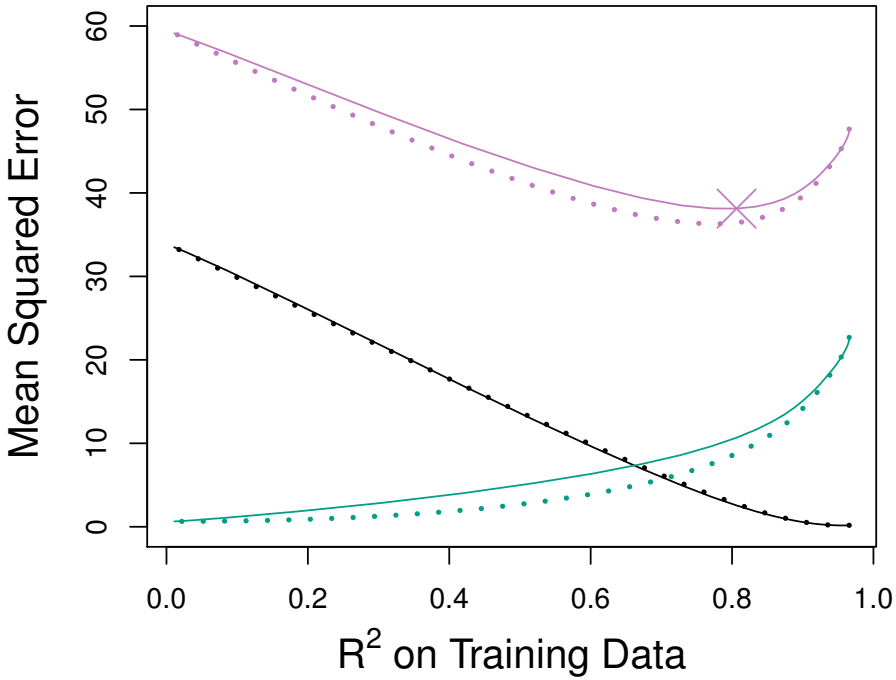
- Limiting case: $q = 0$ counts # nonzero components.

Linear Model Selection and Regularization

Comparison of ridge regression with LASSO

Simulated data using all $p = 45$ predictors: ($\beta_j \neq 0 \; \forall j$ in true model)





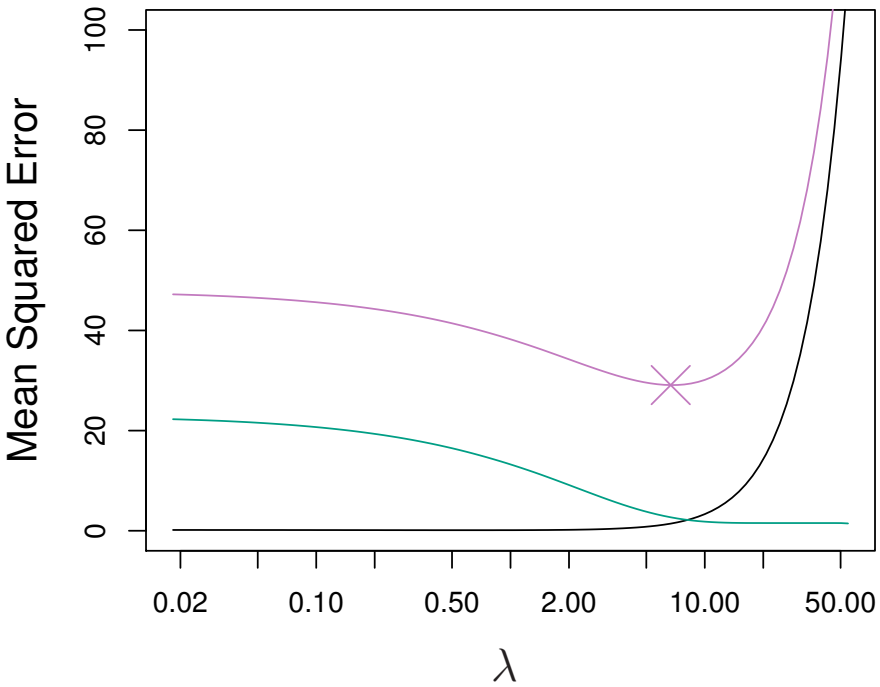
Left: Test MSE (purple), squared bias (black) and variance (green) of LASSO for different values of λ.

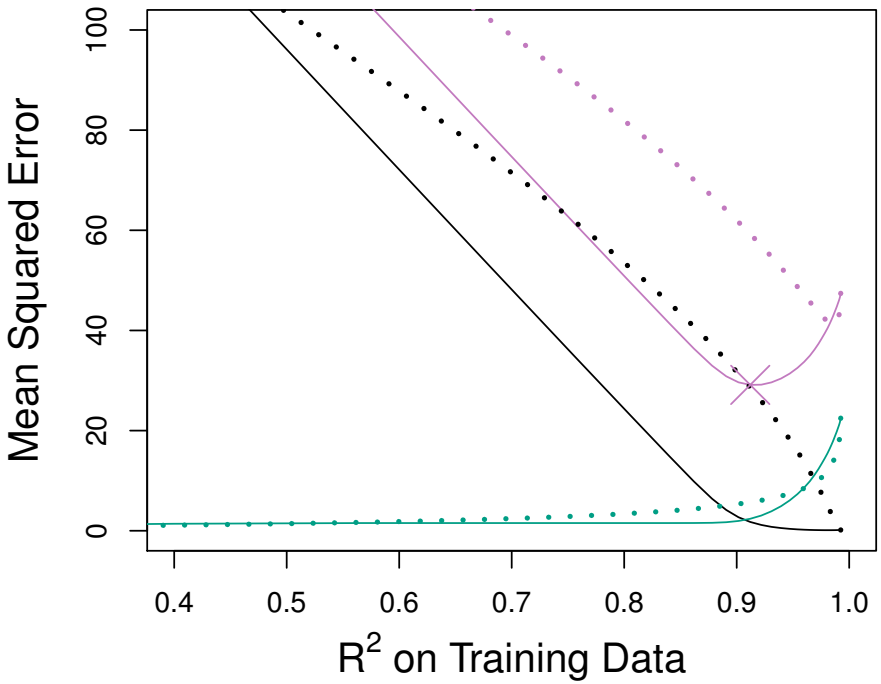
Right: Comparison of test MSE (purple), squared bias (black) and variance (green) against training R^2; dotted lines denote corresponding quantities for ridge regression.

Linear Model Selection and Regularization

Comparison of ridge regression with LASSO

Simulated data using only 2 out of $p = 45$ predictors: (only two $\beta_j \neq 0$ in true model)





Left: Test MSE (purple), squared bias (black) and variance (green) of LASSO for different values of λ.

Right: Comparison of test MSE (purple), squared bias (black) and variance (green) against training R^2; dotted lines denote corresponding quantities for ridge regression.

Assume data matrix $\boldsymbol{X} = \boldsymbol{I}$ $(p = n)$ and $\overline{y} = 0$.
LS problem reduces to minimizing

$$\sum_{j=1}^{p}(y_j - \beta_j)^2, \qquad \text{hence } \beta_j = y_j,\ j = 1, \ldots, p. \tag{6.11}$$

Ridge regression and lasso estimation result from minimizing

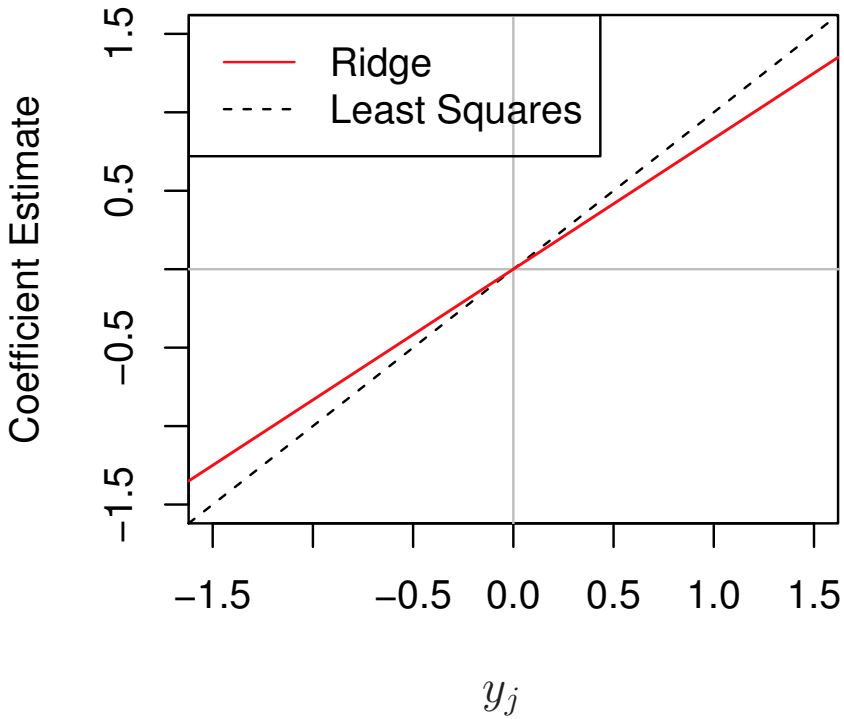
$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad \text{and} \quad \sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

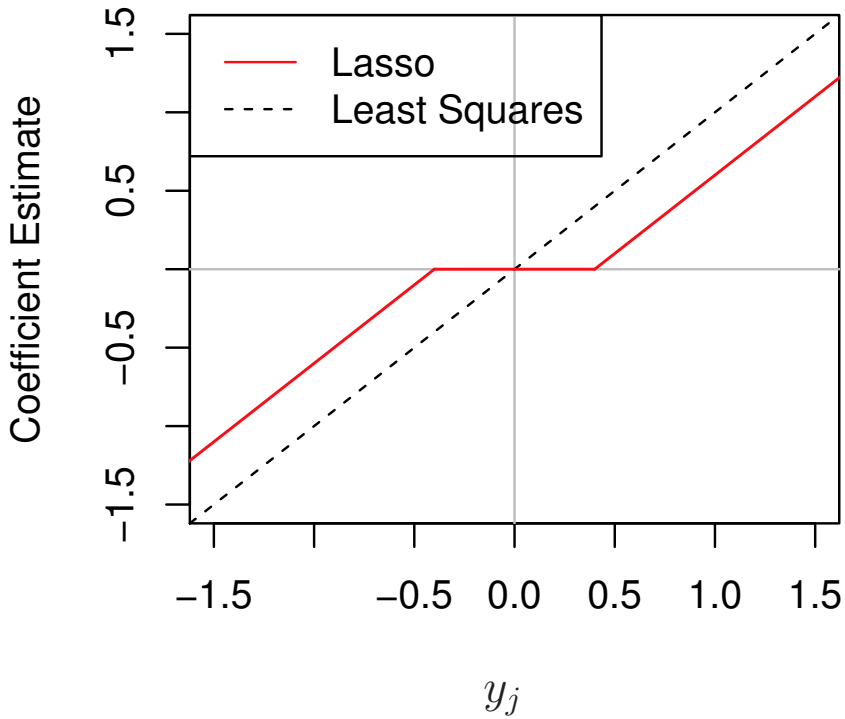
respectively, with solutions

$$\hat{\boldsymbol{\beta}}^R = \frac{y_j}{1+\lambda}, \qquad \hat{\boldsymbol{\beta}}^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2, \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2, \\ 0 & \text{if } |y_j| \le \lambda/2. \end{cases}$$

Linear Model Selection and Regularization

Simple special case for ridge regression and the lasso





Ridge regression (left) and lasso (right) estimates for one variable of special case $\boldsymbol{X} = \boldsymbol{I}$ and $p = n$.

General case: more complicated (of course), but basic mechanism still holds:

- Ridge regression: shrinks every dimension roughly by same proportion.
- Lasso: shrinks all components to zero by similar amount, sufficiently small coefficients damped to zero.

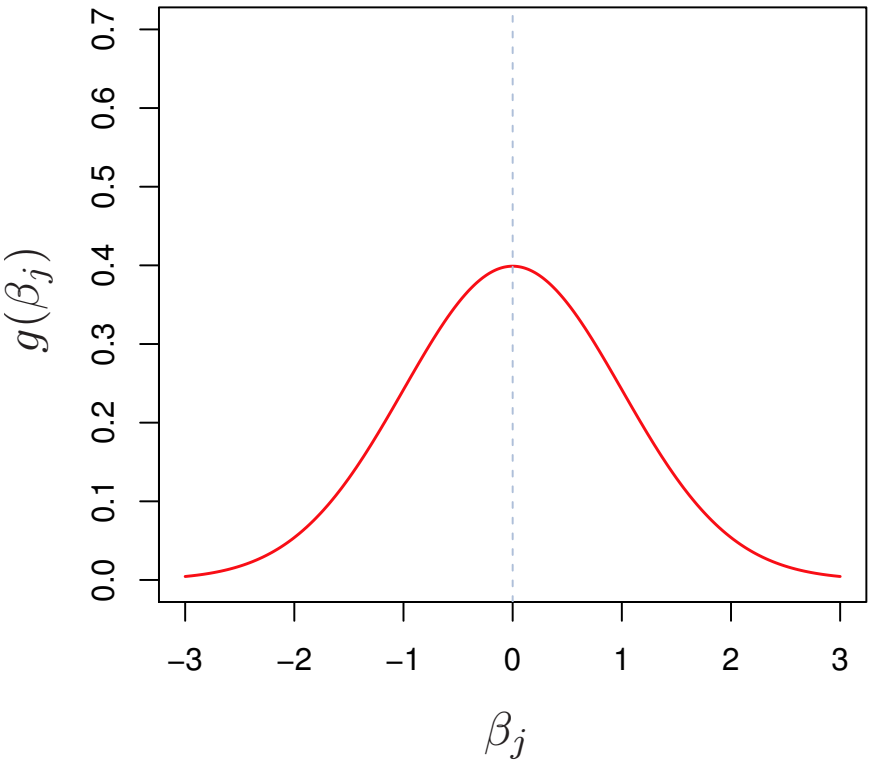
- Assume prior distribution on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, with density $p(\boldsymbol{\beta})$.
- Likelihood of data: $f(Y|X, \boldsymbol{\beta})$, $X = (X_1, \ldots, X_p)$.
- Bayes' rule then says (noting X is fixed)

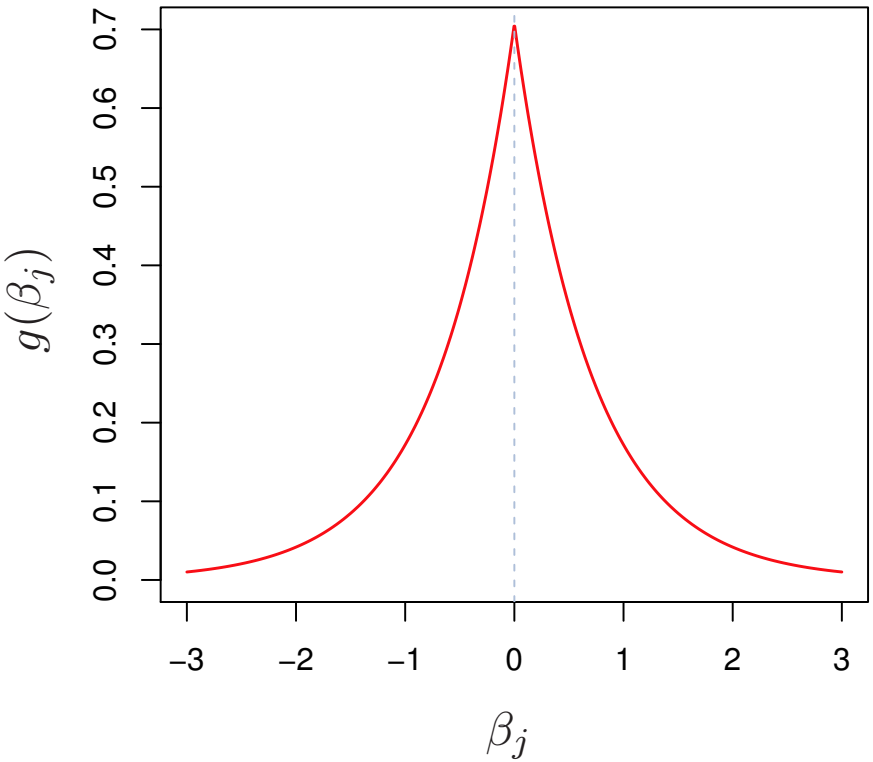
$$p(\boldsymbol{\beta}|X, Y) \propto f(Y|X, \boldsymbol{\beta}) \cdot p(\boldsymbol{\beta}|X) = f(Y|X, \boldsymbol{\beta}) \cdot p(\boldsymbol{\beta}).$$

- Assume standard linear model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$, with independent Gaussian noise and $p(\boldsymbol{\beta}) = \prod_{j=1}^{p} g(\beta_j)$ for pdf g.
- Ridge regression/lasso results from two special cases for g:
  - g centered Gaussian, λ-dependent variance, then ridge regression estimate is **posterior mode** (and posterior mean) of $\boldsymbol{\beta}$.
  - g centered Laplace distribution with λ-dependent scale parameter, then posterior mode for $\boldsymbol{\beta}$ given by lasso estimate. (Not posterior mean; posterior mean itself not sparse.)

Linear Model Selection and Regularization

Bayesian interpretation for ridge regression and the lasso





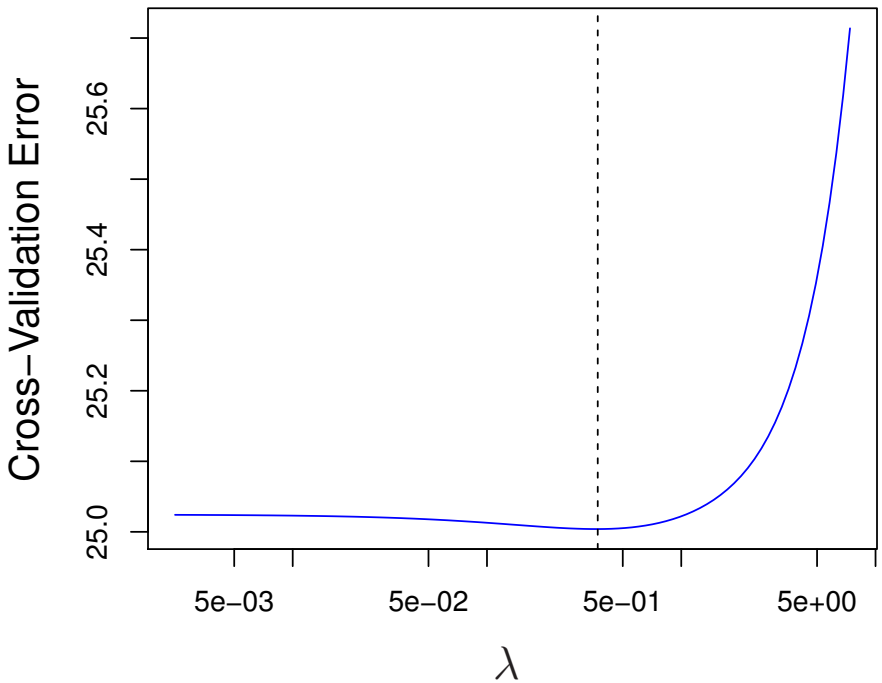
Prior densities for Bayesian interpretation of shrinkage methods.
Left: centered Gaussian prior density, results in posterior distribution with ridge regression solution as posterior mode.

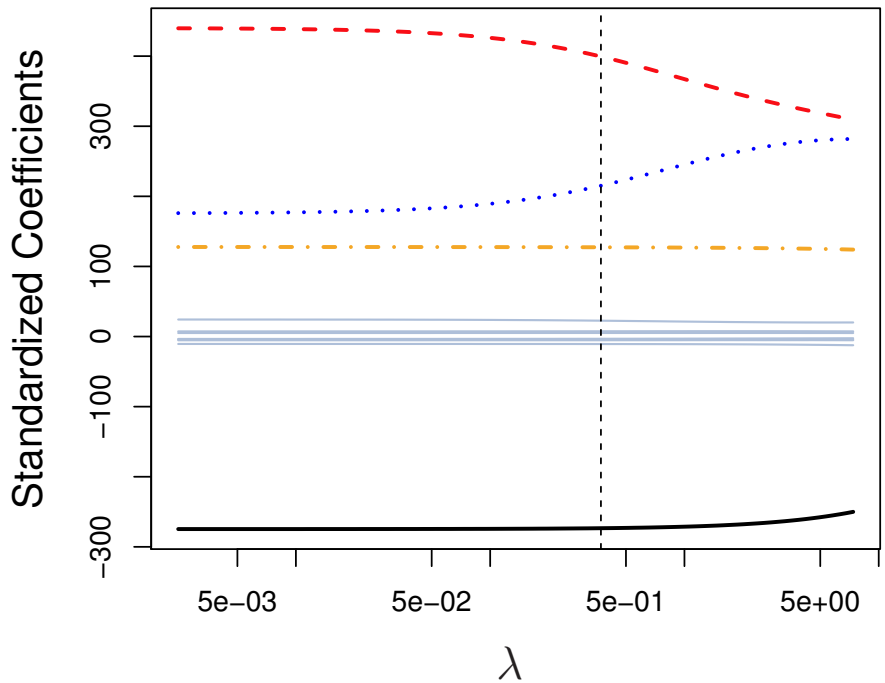
Right: centered Laplace (double-exponential) prior density, results on lasso solution as posterior mode.

Linear Model Selection and Regularization

Selection of λ

- Model selection methods required measure of goodness to compare models.
- Shrinkage methods require selection of shrinkage parameter λ.
- **Cross-validation** approach: fix a grid of λ values; compute cross-validation error for each λ; select λ with smallest error; refit this model with all available observations.



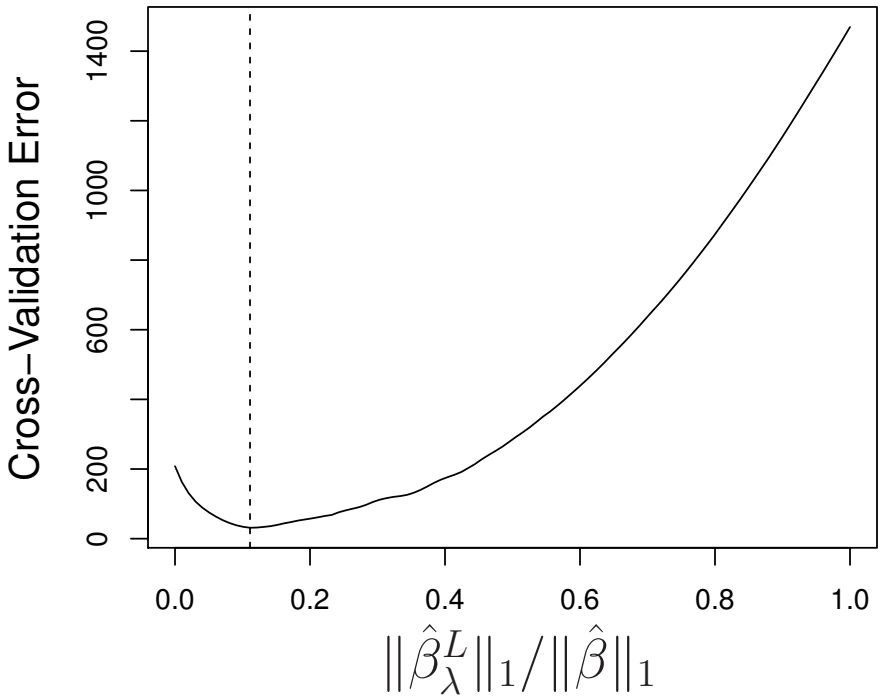


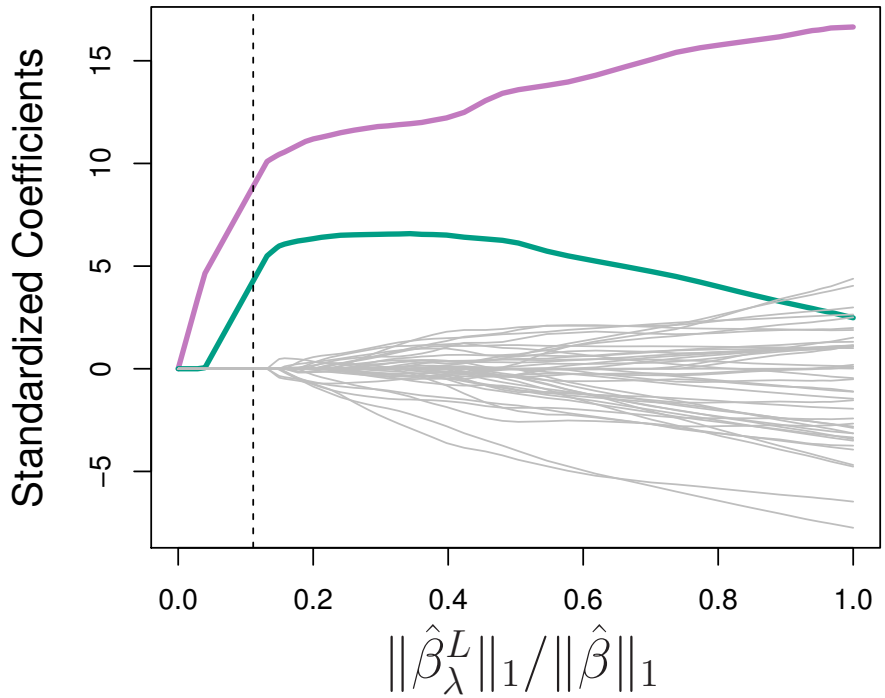
Left: LOOCV errors vs. λ for ridge regression applied to `Credit` data set.

Right: Coefficient estimates vs. λ. Vertical dashed line indicates selected λ.

Linear Model Selection and Regularization

Selection of λ





10-fold CV applied to data set from Slide 281.

Left: CV error. Right: coefficient estimates. Vertical dashed line indicates CV error-minimizing λ. Colored lines represent 2 predictors related to response, grey lines unrelated predictors (**signal** vs. **noise**).

Lasso assigns relevant predictors much larger estimates; CV chooses λ for which irrelevant predictors set to zero. Compare LS estimate (far right).

Contents

- Up to now: control variance by removing predictor variables or shrinking coefficients.
- Now: reduce variance by projecting into subspace of dimension $M < p$.
- Set

$$Z_m := \sum_{j=1}^{M} \phi_{j,m} X_j, \ m = 1, \ldots, M, \qquad \text{i.e.,} \ \ \boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Phi}, \ \boldsymbol{\Phi} \in \mathbb{R}^{p \times M}. \tag{6.12}$$

- Fit standard linear regression model

$$Y = \theta_0 + \theta_1 Z_1 + \cdots + \theta_M \beta_M + \varepsilon. \tag{6.13}$$

- **Dimension reduction**: fit $M + 1 < p + 1$ coefficients.
- For well-chosen $\boldsymbol{\Phi}$, this reduced-dimension approach can outperform LS.

- Note:

$$\sum_{m=1}^{M} z_{i,m}\theta_m = \sum_{m=1}^{M}\sum_{j=1}^{p} \phi_{j,m}x_{i,j}\theta_m = \sum_{j=1}^{p} x_{i,j}\sum_{m=1}^{M} \phi_{j,m}\theta_m = \sum_{j=1}^{p} x_{i,j}\beta_j$$

  with $\beta_j := \sum_{m=1}^{M} \phi_{jk,m}\theta_m$.

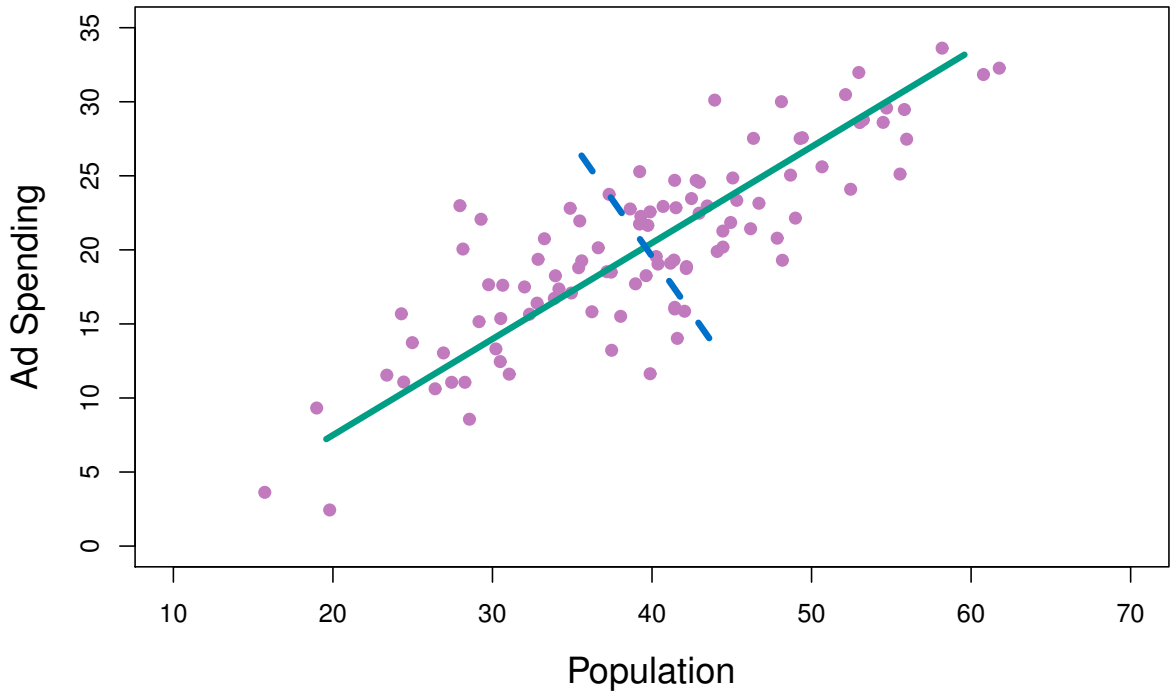
- In matrix terms:

$$\boldsymbol{Z\theta} = \boldsymbol{X\Phi\theta} =: \boldsymbol{X\beta}, \qquad \boldsymbol{\beta} := \boldsymbol{\Phi\theta}$$

  Hence can view (6.13) as special case of original linear model (6.1).

- Dimension reduction constrains $\boldsymbol{\beta}$ by making it a linear function of $M < p$ variables $\{\theta_m\}_{m=1}^{M}$.
- May introduce bias, but when $p \gg n$ this is outweighed by resulting variance reduction.
- Next: 2 ways of choosing $\boldsymbol{\Phi}$.

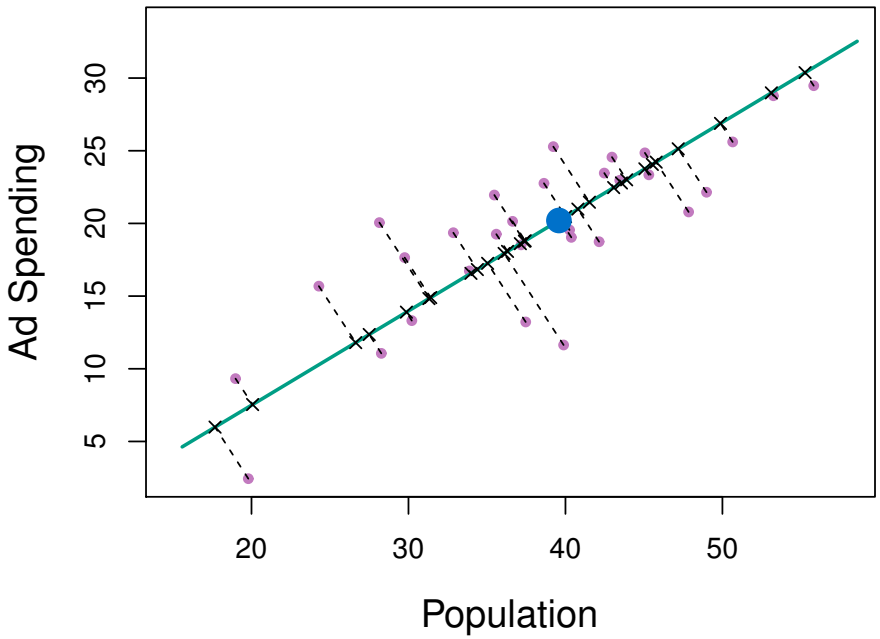
- **Principal components analysis** (PCA): approach for deriving a low-dimensional feature set from a large set of variables.
- First principal component: direction in $\mathbb{R}^p$ in which observations vary the most.

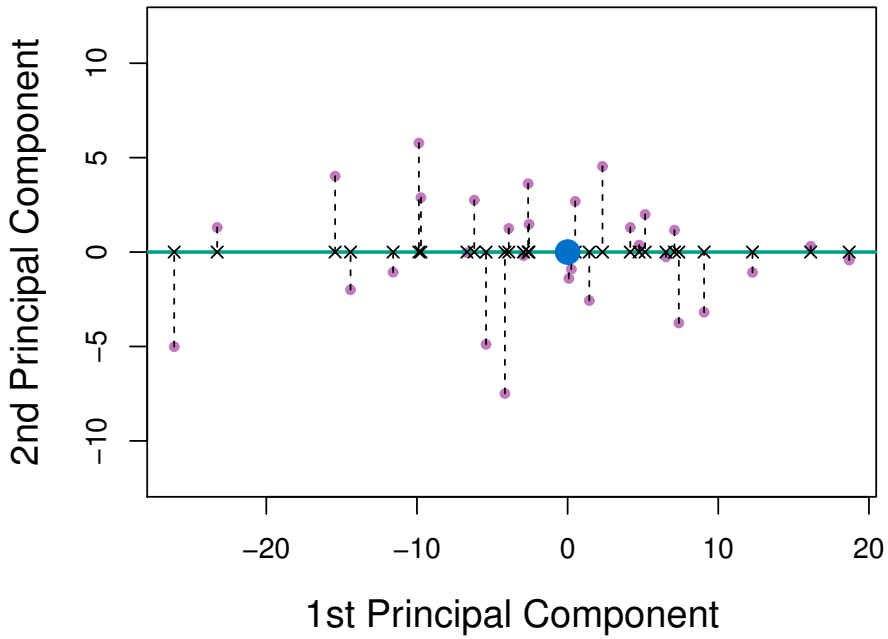


Population size `pop` vs. ad spending `ad` for 100 cities (purple dots). Green solid line: first principal component; blue dashed line: second principal component.

Linear Model Selection and Regularization

Principal components regression





- Project data on direction (line) along which it varies most.
- For pop / ad data: $\phi_{1,1} = 0.839$, $\phi_{2,1} = 0.544$, giving

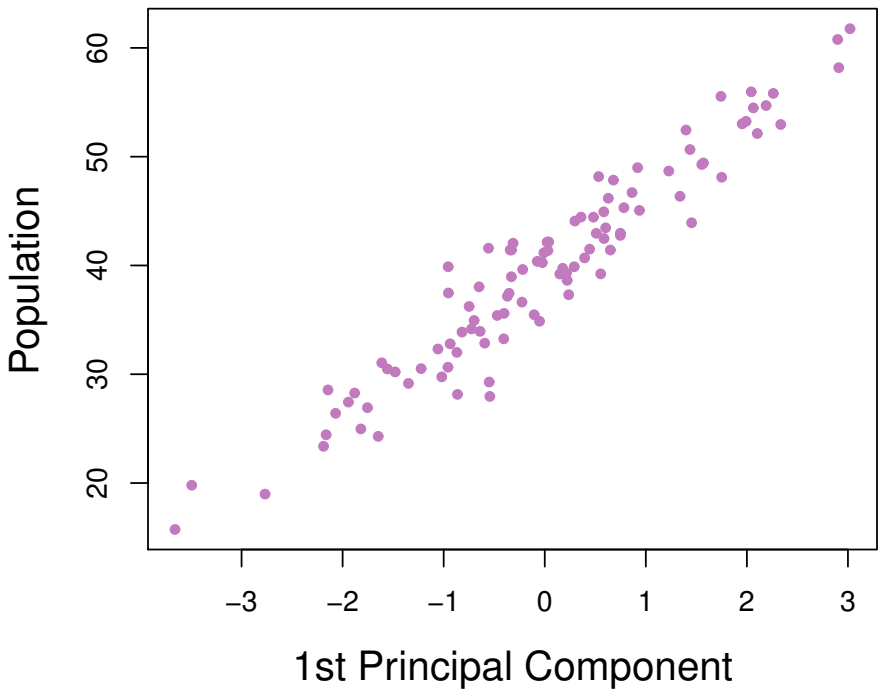
$$Z_1 = 0.839 \times (\mathtt{pop} - \overline{\mathtt{pop}}) + 0.544 \times (\mathtt{ad} - \overline{\mathtt{ad}})$$

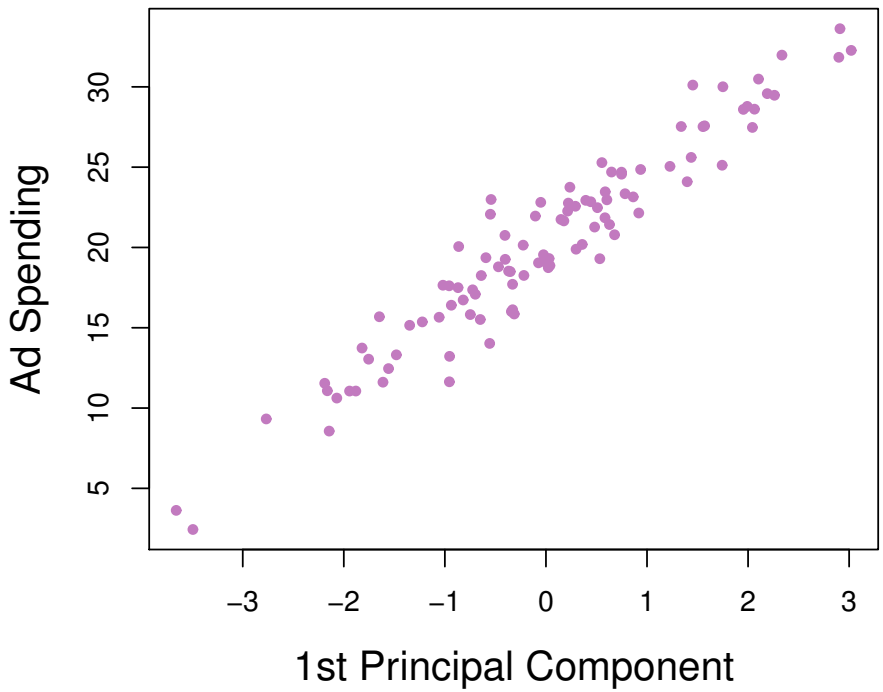
- Out of every (normalized) linear compbination of the (centered) variables, Z_1 has maximal variance.

- Principal component data vector ("**scores**") has same length n, e.g.

$$z_{i,1} = 0.839 \times (\mathtt{pop}_i - \overline{\mathtt{pop}}) + 0.544 \times (\mathtt{ad}_i - \overline{\mathtt{ad}}), \qquad i = 1, \ldots, n.$$

- Alternative interpretation of PCA: 1st principal component vector defines line as close as possible to data in sense of minimizing sum of squared perpendicular distances between each data point and this line.
-



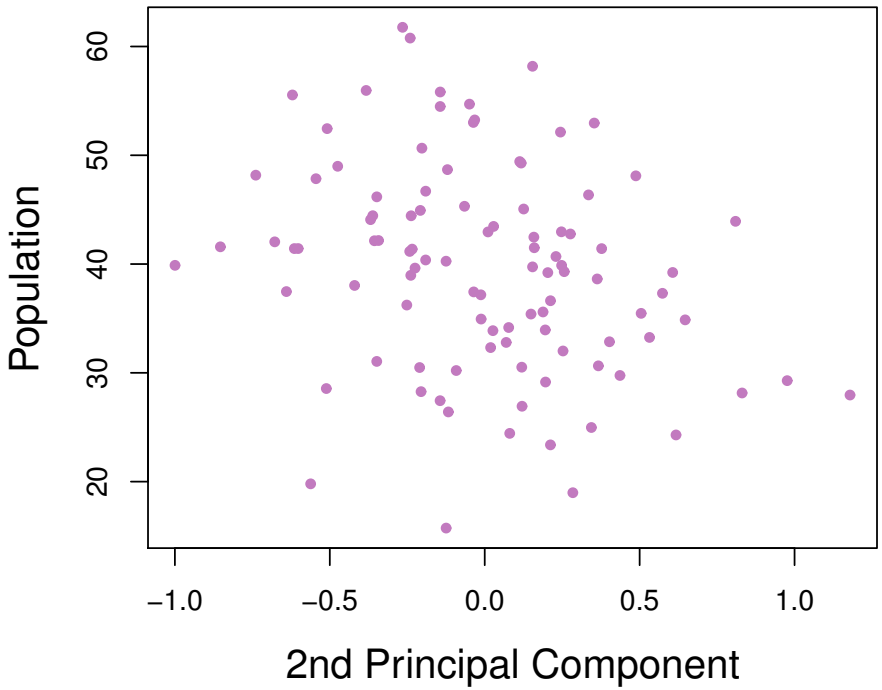


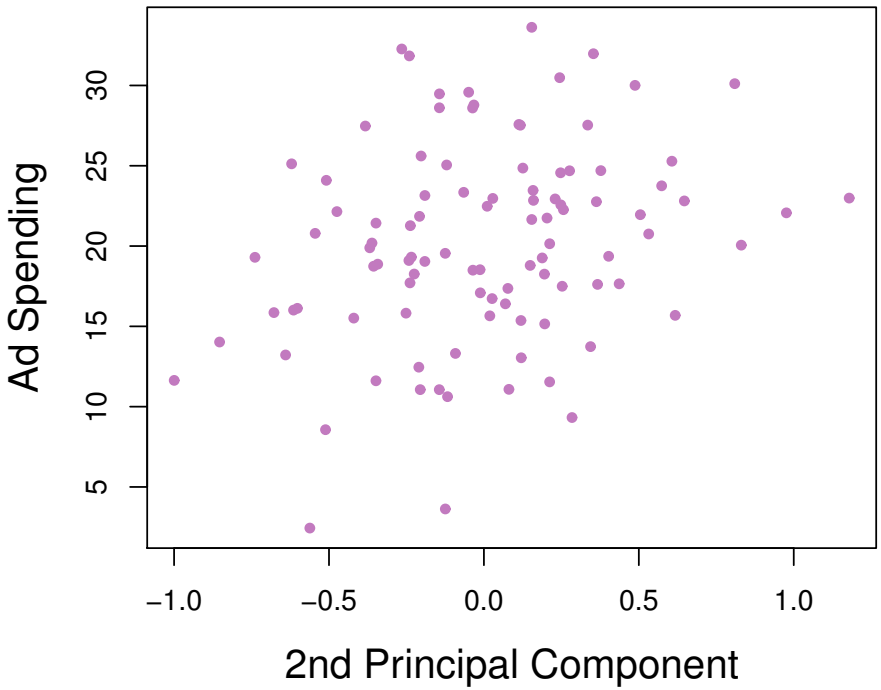
First principal components scores $z_{i,1}$ for `pop` and `ad`. Strong relationship in both cases, i.e., principal component captures most of the information contained in the two predictors.

- Second principal component Z_2: direction of largest variance among all linear combinations of predictor variables which is *orthogonal* to (uncorrelated with) Z_1.
- Here:
$$Z_2 = 0.544 \times (\texttt{pop} - \overline{\texttt{pop}}) - 0.839 \times (\texttt{ad} - \overline{\texttt{ad}}).$$
Since $p = 2$, this covers all of remaining variance.
- Of these, Z_1 contains most of the information, cf. much larger variation in Z_1-coordinate than Z_2-coordinate in right panel of figure on Slide 292.
- Plot on Slide 296 displays $z_{i,2}$ against `pop` and `ad` predictors: much less relationship than with Z_1.
Thus, Z_1 sufficient to explain most of variability in data set.
- For p predictor variables, can construct up to p principal components.

Linear Model Selection and Regularization

Principal components regression

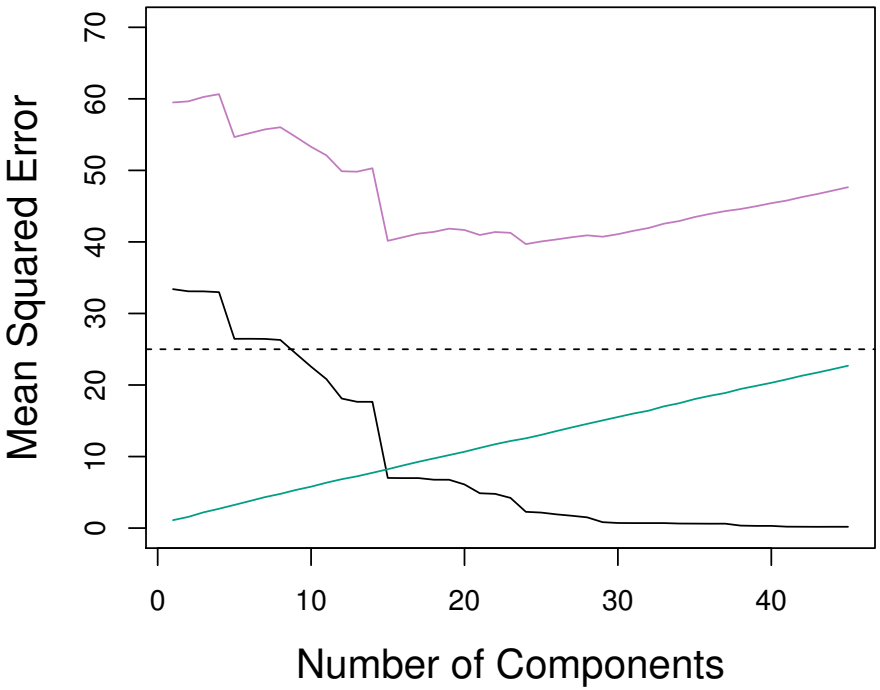


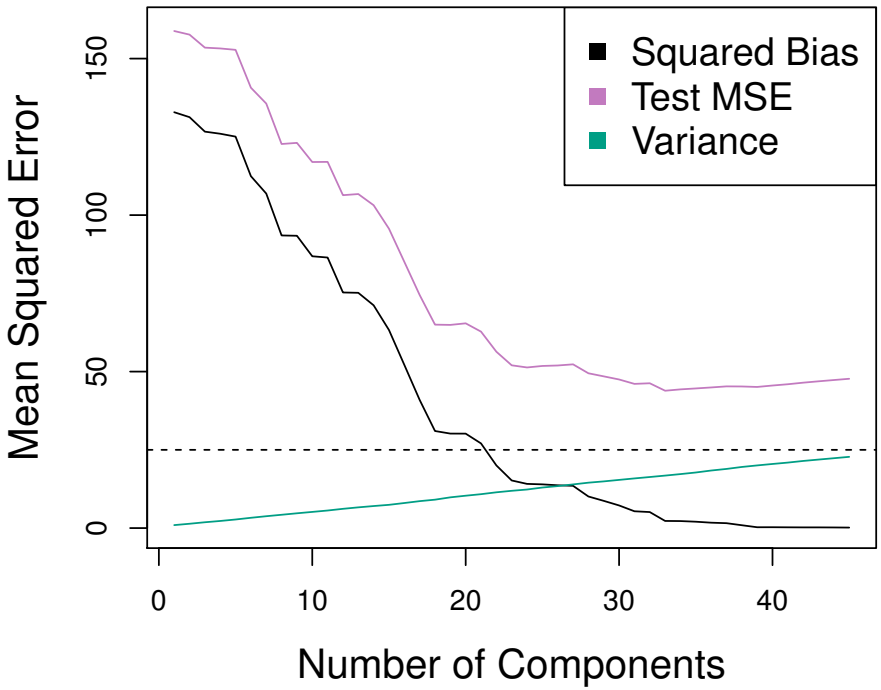


- **Principal components regression** (PCR): construct first M proincipal components $Z_1, \ldots, Z_M$, use these in a linear regression model fit by LS.
- Guiding principle: directions in span of $X_1, \ldots, X_p$ with most variance are the directions associated with response Y.
- Under this assumption, fitting LS model to $Z_1, \ldots, Z_M$ will yield better predictions than fitting $X_1, \ldots, X_p$, since most information related to response Y contained in $Z_1, \ldots, Z_M$, and estimating $M \ll p$ coefficients avoids overfitting.

Linear Model Selection and Regularization

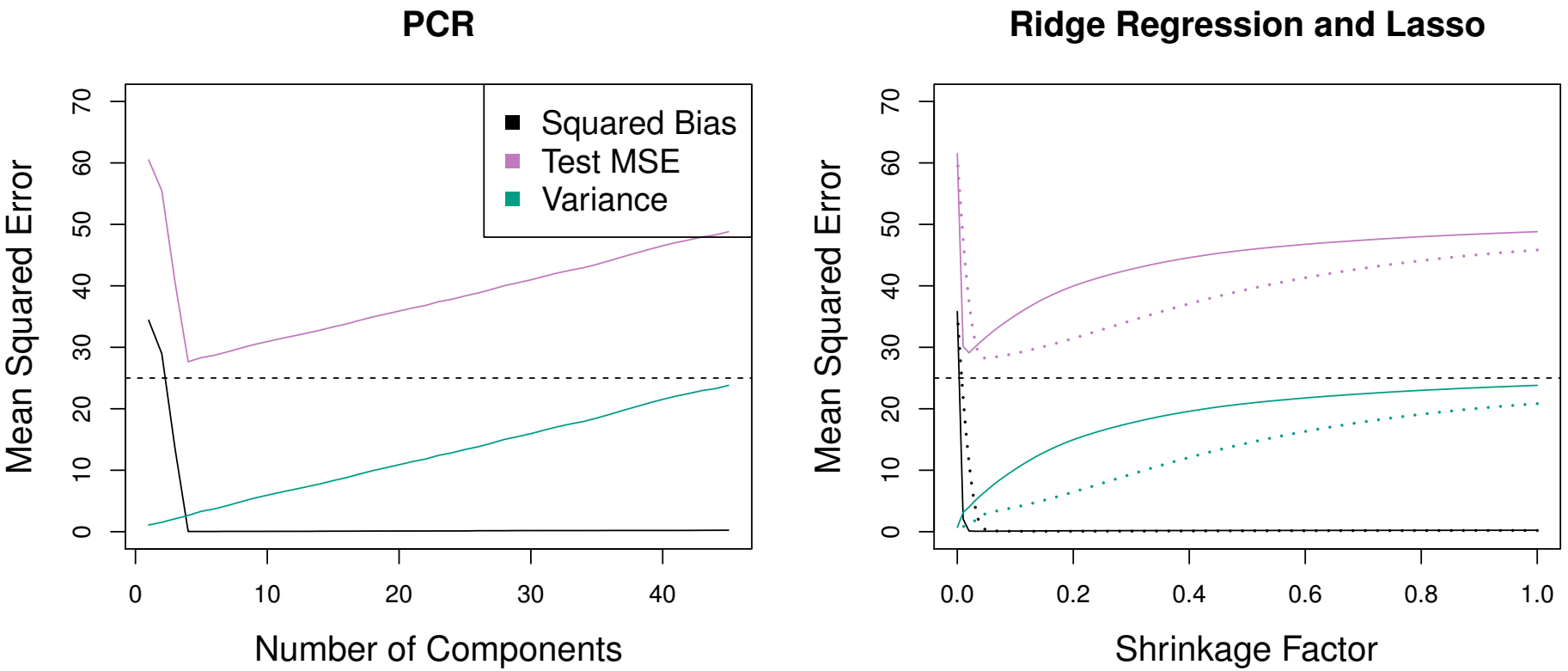
Principal components regression





PCR fits to data sets from Slide 280 (left) and Slide 281 (right): MSE against # proncipal components M. More components reduces bias, increases variance (U-shape). $p = M$ coincides with LS fit of original predictors. Compared with ridge regression and lasso results in figures on Slides 272, 280 and 281, PCR seen to underperform shrinkage.

Worse performance of PCR in previous example due to fact that many principal components needed needed to explain response.



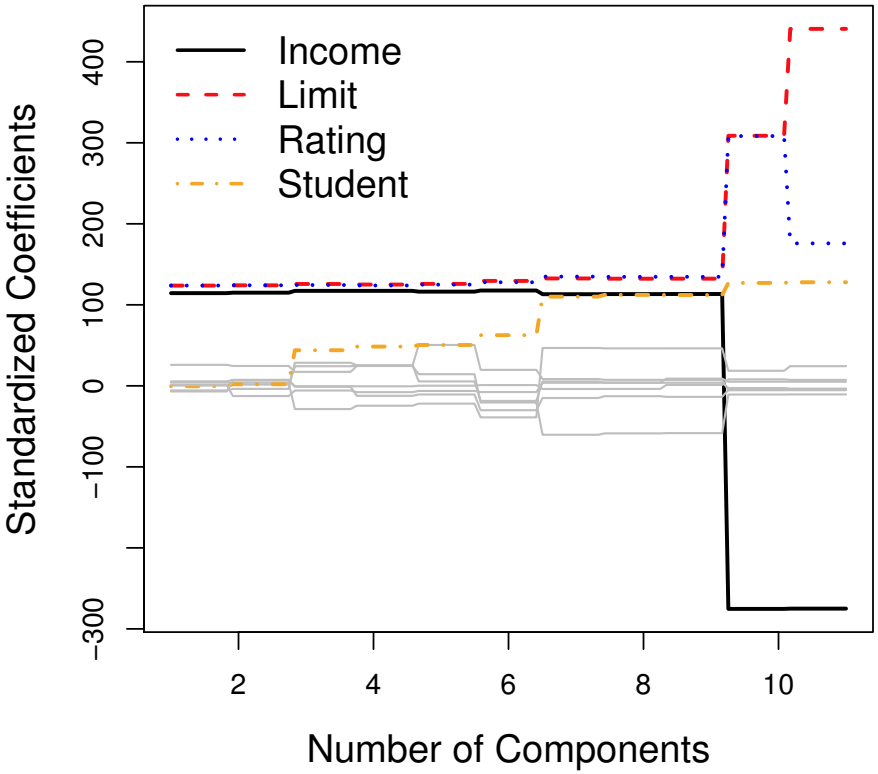
Data generated in such a way that response depends exclusively on first 5 principal components. Left: PCR, MSE has clear minumum at $M = 5$.

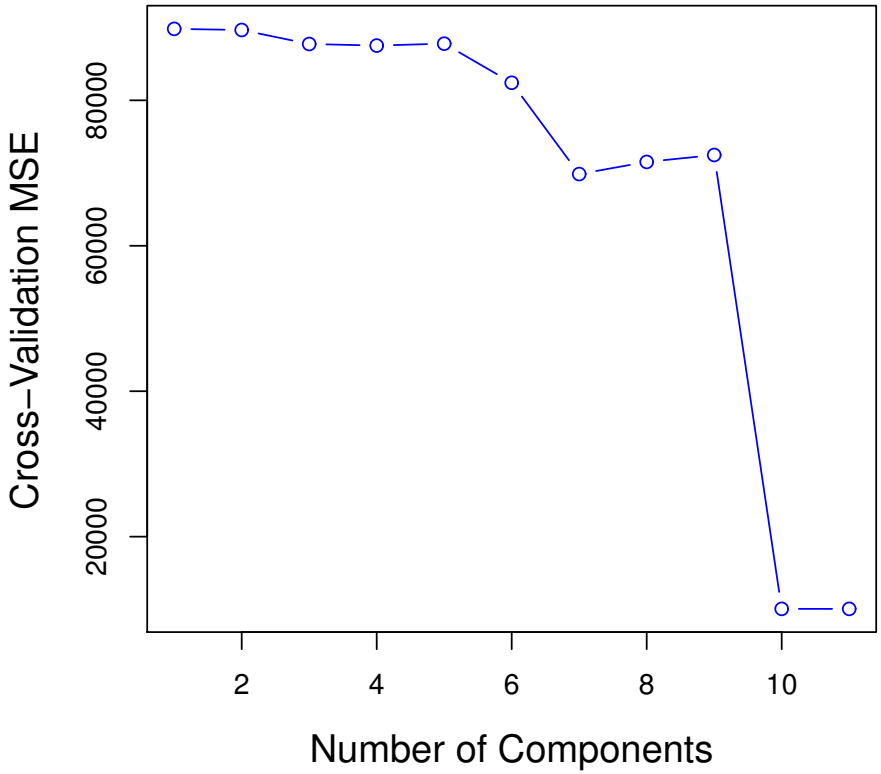
Right: ridge regression (dotted) and lasso (solid) results.

- PCR uses $M < p$ new variables, but these all still depend on original predictors.
- Hence, PCR not a feature selection method.
- In this aspect, PCR closer to ridge regression than lasso.
- Ridge regression can be viewed as a continuous version of PCR.
- # principal components M can be chosen by CV.
- Recommended: first standardize data.

Linear Model Selection and Regularization

Principal components regression



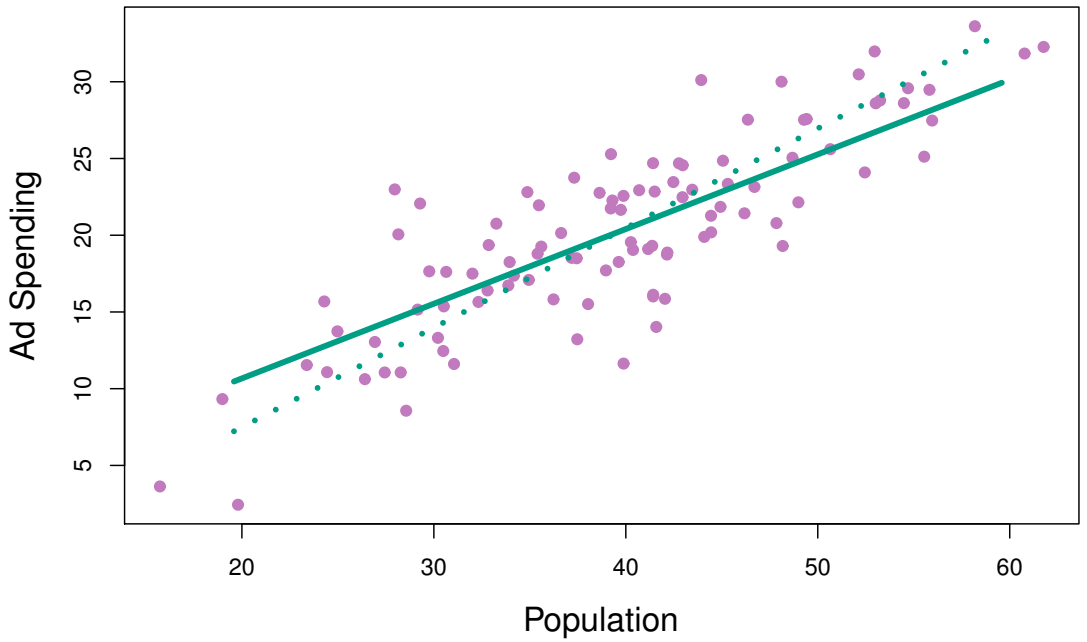


PCR applied to `Credit` data set.

Left: standardized coefficients. Right: CV MSE against M.

Lowest error for 10 components (only one less than full model).

- PCR only looks at predictor variability, not at response.
- In this sense, it is **unsupervised**.
- **Partial least squares** (PLS): **supervised** variant of PCR: find linear combination of predictors containing most variability and best explain response.
- To construct Z_1, set each coefficient in $Z_1 = \sum_{j=1}^{p} \phi_{j,1} X_j$ to coefficient of simple linear regression of Y onto X_j. Results in coefficient proportional to $\mathsf{Cor}(X_j, Y)$.
  This places highest weight on variables most strongly related to response Y.
- To identify Z_2, first adjust all predictors for Z_1 by regressing these on Z_1 and taking residuals. Interpretation: remaining information not explained by first PLS direction. Compute Z_2 using this orthogonalized data just as Z_1 was computed using original data.
- In the same way, compute further PLS directions $Z_3, \ldots, Z_M$.

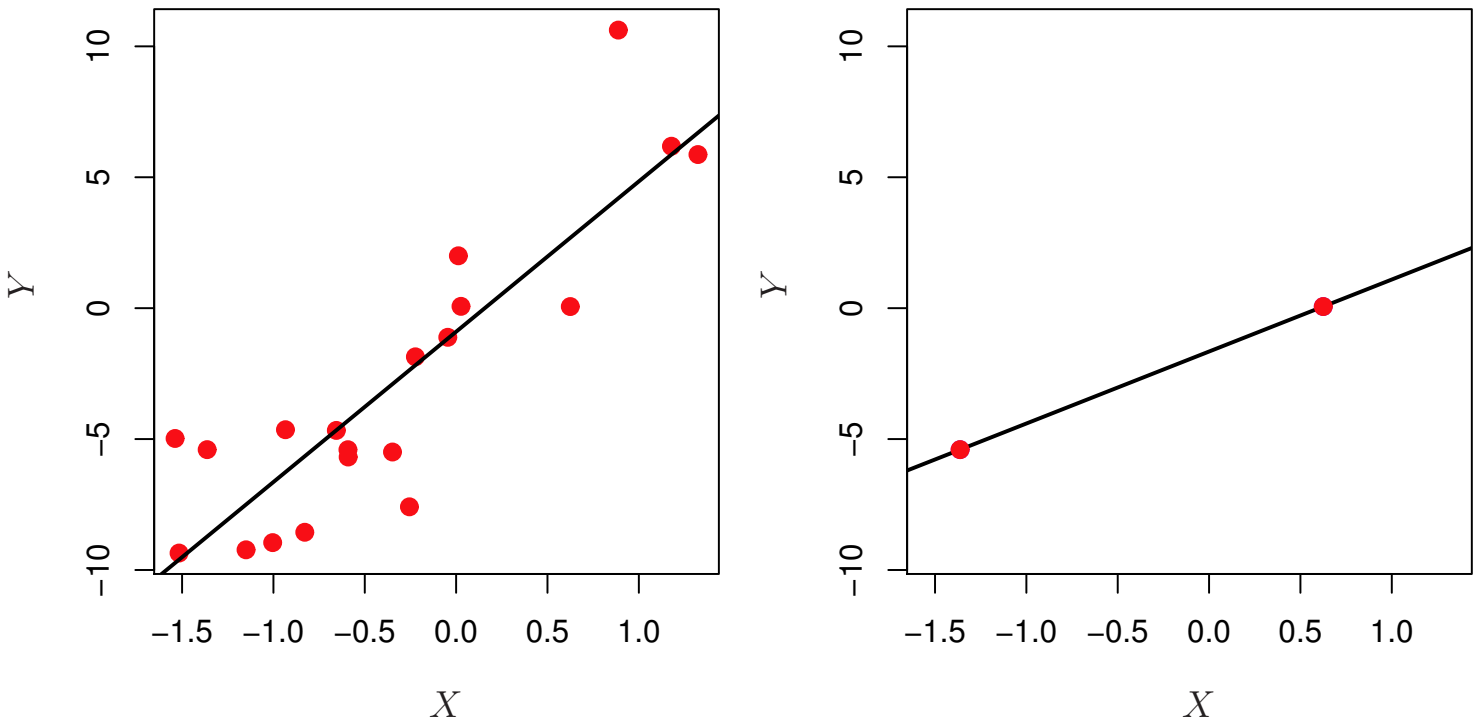


PLS on synthetic data set giving `Sales` data in each of 100 regions as response to two predictors `Population Size` and `Advertising Spending`. Solid line: first PLS direction, dotted: first principal components direction.

Contents

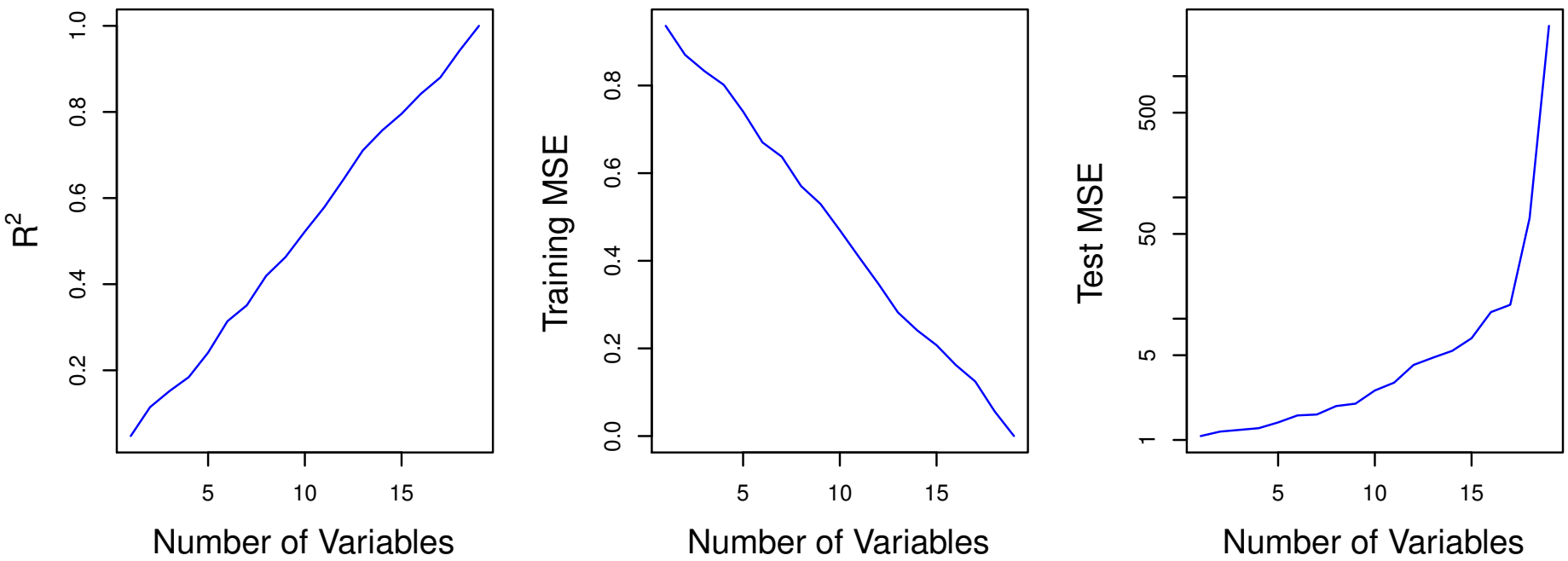
- Most traditional statistical techniques: $n \gg p$ (**low-dimensional setting**).
- Typical example: Predict patient's blood pressure based on age, gender, body mass index (BMI). Three predictors, and typically thousands of patients' data.
- More recently, in many fields such as medicine, finance, marketing, trend towards collecting almost unlimited number of feature measurements (p large), while cost of obtaining sufficienly many samples prohibitive.
- Example: in place of age, gender, BMI, collect measurements of half million **single nucleotide polymorphisms**, i.e., common individual DNA mutations. Results in $p \approx 500,000$, $n \approx 200$.
- Example: 'Bag-of-words' model to understand customers' online shopping patterns, using as features all search terms entered in search engine (binary feature vector). Only few hundred users consented to their data being used. Results in $n \approx 100$, p much larger.

- When $p \geq n$ LS cannot (should not) be used, since data will be fit perfectly.
- Example: $p = 1$ $n = 20$ vs. $n = 2$:



Right model will not generalize well (overfitting), model too flexible..

Another example: $n = 20$ observations for $1 \leq p \leq 20$ features, each *completely unrelated* to response.



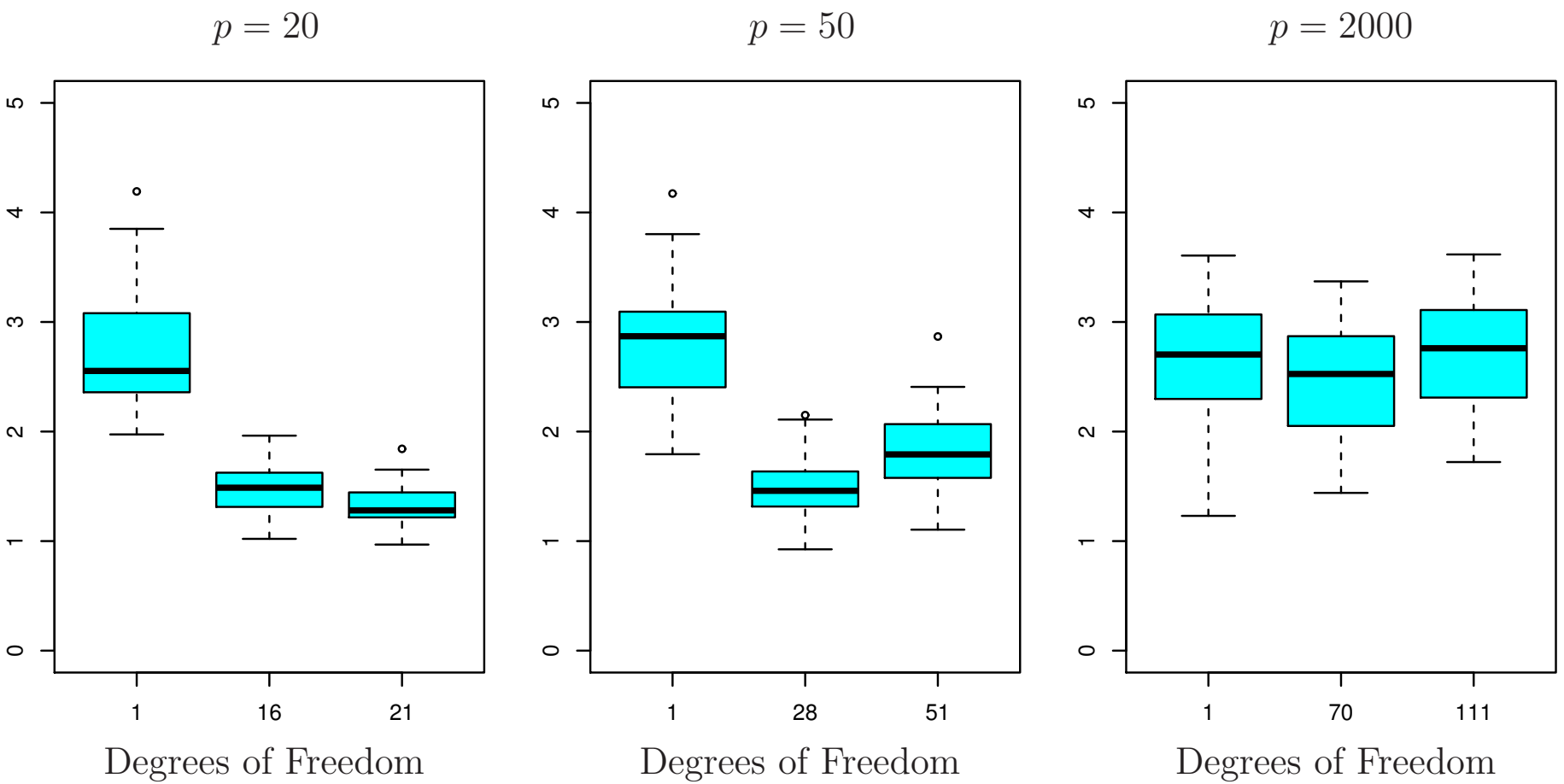
As p increases, $R^2 \to 1$, training MSE $\to 0$ despite no relation of predictors to response. At the same time, test MSE sharply increases as model increasingly flexible.

Casual observer may find large model superior if only first two quantities monitored.

- Model selection techniques based on C_p, AIC, BIC not appropriate for high-dimensional setting, as estimating $\hat{\sigma}^2$ problematic.
- Adjusted R^2 may easily yield value of 1 in high-dimensional setting.
- Less flexible regression models (stepwise selection, shrinkage, PCR) particularly useful in high dimensions. Avoid overfitting by constraining flexibility.
- Next figure: Lasso on $n = 100$ simulated training observations using $p = 20, 50$ and $2,000$ features, of which 20 related to response.
  Then MSE evaluated on independent test set.

Linear Model Selection and Regularization

The high-dimensional setting: what goes wrong?



- For $p = 20$, lowest test MSE for low value of λ. For larger p, best model obtained for larger λ. When $p = 2000$, lasso performs badly for all values of λ.
- Rather than λ, plot shows **degrees of freedom** of model, i.e., # nonzero coefficients of lasso estimate.

Summary:

1. Shrinkage plays key role in high dimensions.
2. Correct value of tuning parameter essential.
3. Test error increases with dimension, unless additional features informative.

- Third observation related to **curse of dimensionality**: quality of model need not increase as features added.
- Compare left and right panel in figure: test MSE almost doubles as p increased from 20 to 2000.
- **Noise features** (not related to response) increase dimension, exacerbate overfitting danger.
- Adding features truly related to response will generally improve model.
- New sensor technology allowing for millions of observations can lead to worse results if features not relevant. Even if relevant, variance incurred by fitting their coefficients may outweigh reduction in bias from additional features.

Linear Model Selection and Regularization

The high-dimensional setting: what goes wrong?

- In high dimensions: **collinearity** problem extreme. (Why?)
- Never know which variables truly predictive, can never obtain best coefficients.
- At best: assign large coefficients to variables correlated with variables truly predictive for response.
- For $p > n$ can easily obtain useless model with zero residual.
- Traditional measures of model quality based on training data often highly misleading in high dimensions.
- Reporting MSE on independent test data particularly important here.

Contents

Theorem 6.1 (Gauss-Markov theorem)

Given observations $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, $i = 1, \ldots, n$, for which the uncorrelated random noise variables ε_i have mean zero and constant variance $\sigma^2 > 0$, and assuming that the observation vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in \mathbb{R}^n$ are linearly independent, then the least squares estimate

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \quad \boldsymbol{X} = [\boldsymbol{x}_1 | \cdots | \boldsymbol{x}_p] \in \mathbb{R}^{n \times p}, \quad \boldsymbol{y} \in \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

has minimal variance among all linear unbiased estimators of $\boldsymbol{\beta}$.

[7] C.F. Gauss, 1777–1855, A.A. Markov, 1856–1922

Optimality of LS Estimate

Remarks

- No assumption is made on the distribution of the errors, only on their first two moments.
- The theorem also holds if $\mathbf{Var}\,\varepsilon = \boldsymbol{\Sigma}$ is a (nonsingular) covariance matrix. In this case the best linear unbiased estimator solves the **weighted least squares** problem

$$\|y - X\beta\|_{\boldsymbol{\Sigma}} \to \min_{\beta}, \qquad \|x\|_{\boldsymbol{\Sigma}}^2 = x^\top \boldsymbol{\Sigma}^{-1} x.$$

- The theorem does not say there are no better estimators than LS, only that any better estimators are either nonlinear or biased.
- Examples of biased estimators are ridge regression and the lasso.
- ESL:
  "Most models are distortions of the truth, and hence are biased; picking the right model amounts to creating the right balance between bias and variance.

Theorem 6.2 (Singular value decomposition)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ of rank r, there exist orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ as well as a "diagonal" matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_r & O \\ O & O \end{bmatrix} \in \mathbb{R}^{n\times p} \qquad \text{where} \qquad \boldsymbol{\Sigma}_r = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$$

and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, such that

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top. \tag{SVD}$$

- The positive numbers $\sigma_1, \ldots, \sigma_r$ are called the **singular values** of $\boldsymbol{A}$.
- The columns of $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p]$ are the **left and right singular vectors**, respectively.

1. Representation of $\boldsymbol{A}$ as sum of rank-1 matrices:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r]\, \boldsymbol{\Sigma}_r\, [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r]^\top = \sum_{k=1}^{r} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$$

2. Singular vector mapping properties:

$$\boldsymbol{A}\boldsymbol{v}_k = \begin{cases} \sigma_i \boldsymbol{u}_i & k = 1, 2, \ldots, r, \\ \boldsymbol{0} & k = r+1, \ldots, p \end{cases}$$

and

$$\boldsymbol{A}^\top \boldsymbol{u}_k = \begin{cases} \sigma_k \boldsymbol{v}_k & k = 1, 2, \ldots, r, \\ \boldsymbol{0} & k = r+1, \ldots, n. \end{cases}$$

3.

$\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$	is an ON-basis of	$\mathscr{R}(\boldsymbol{A})$.
$\{\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_m\}$	is an ON-basis of	$\mathscr{N}(\boldsymbol{A}^\top) = \mathscr{R}(\boldsymbol{A})^\perp$.
$\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\}$	is an ON-basis of	$\mathscr{R}(\boldsymbol{A}^\top) = \mathscr{N}(\boldsymbol{A})^\perp$.
$\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n\}$	is an ON-basis of	$\mathscr{N}(\boldsymbol{A})$.

4. Eigenspaces of $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$:
   - $\sigma_1^2, \ldots, \sigma_r^2$ are the non-zero eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$, respectively:

$$\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\begin{bmatrix}\boldsymbol{\Sigma}_r^2 & o\\ o & o\end{bmatrix}\boldsymbol{V}^\top,$$

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top = \boldsymbol{U}\begin{bmatrix}\boldsymbol{\Sigma}_r^2 & o\\ o & o\end{bmatrix}\boldsymbol{U}^\top.$$

   - In particular, the singular values $\sigma_1, \ldots, \sigma_r$ are uniquely determined by $\boldsymbol{A}$.
   - The right singular vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ form an ON-basis of $\mathbb{R}^p$ of eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$:

$$\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}_k = \begin{cases}\sigma_k^2\boldsymbol{v}_k & k = 1, 2, \ldots, r,\\ \boldsymbol{0} & k = r+1, \ldots, p.\end{cases}$$

     The left singular vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ form an ON-basis of $\mathbb{R}^n$ of eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$:

$$\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u}_k = \begin{cases}\sigma_i^2\boldsymbol{u}_k & k = 1, 2, \ldots, r,\\ \boldsymbol{0} & k = r+1, \ldots, n.\end{cases}$$

Singular Value Decomposition

Properties

5. If $\boldsymbol{A} = \boldsymbol{A}^\top \in \mathbb{R}^{n\times n}$ with non-zero eigenvalues

$$\lambda_1, \ldots, \lambda_r, \qquad |\lambda_1| \geq \cdots \geq |\lambda_r| > 0,$$

   then the singular values of $\boldsymbol{A}$ are given by $\sigma_k = |\lambda_k|$.

6. The (p-dimensional) unit sphere is mapped by $\boldsymbol{A}$ to an ellipsoid (in $\mathbb{R}^n$) with center $\mathbf{0}$ and semi-axes $\sigma_k \boldsymbol{u}_k$ ($\sigma_k := 0$ für $k > r$).

7. For $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ there holds $\|A\|_2 = \sigma_1$ and $\|\boldsymbol{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$.
   For $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ invertible, there holds in addition that $\|\boldsymbol{A}^{-1}\|_2 = \sigma_n^{-1}$.

8. Anlalogous statements hold for complex-valued matrices $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$ ($\boldsymbol{U}, \boldsymbol{V}$ unitary). (In (5) replace $\boldsymbol{A} = \boldsymbol{A}^\top$ by '$\boldsymbol{A}$ normal'.)

Singular Value Decomposition

Best rank-k approximation

- E. Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen. 1. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener. Math. Ann., 63 (1907), pp. 433–476
- C. Eckart, G. Young. The approximation of one matrix by another of lower rank. Psychometrika, 1 (1936), pp. 211–218
- L. Mirsky. Symmetric gauge functions and unitarily invariant norms. Quart. J. Math. Oxford, 11 (1960), pp. 50–59

Theorem 6.3 (Best approximation by matrices of lower rank)

For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ of rank r with SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ the best approximation problem

$$\min\{\|\boldsymbol{A}-\boldsymbol{B}\|_2 \,:\, \boldsymbol{B}\in\mathbb{R}^{n\times p} \text{ and } \operatorname{rank}(B)\le k\}$$

for $k < r$ is solved by

$$\boldsymbol{A}_k := \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \qquad \text{with} \qquad \|\boldsymbol{A}-\boldsymbol{A}_k\|_2 = \sigma_{k+1}.$$

- $\boldsymbol{A}_k$ as above is also the closest rank-k matrix to $\boldsymbol{A}$ in the Frobenius-norm $\|\cdot\|_F$, with distance $\|\boldsymbol{A}-\boldsymbol{A}_k\|_F = \sqrt{\sigma_{k+1}^2+\cdots+\sigma_r^2}$.

- Recall the ridge regression estimate $\hat{\boldsymbol{\beta}}_R$ for the LS problem $\boldsymbol{X}\boldsymbol{\beta} \approx \boldsymbol{y}$ with data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and observation vector $\boldsymbol{y} \in \mathbb{R}^n$: for a given value of the tuning (or regularization) parameter $\lambda \geq 0$ it was defined by

$$\hat{\boldsymbol{\beta}}_R = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, Q_\lambda(\boldsymbol{\beta}), \qquad Q_\lambda(\boldsymbol{\beta}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

- Rewriting the objective function $Q_\lambda(\boldsymbol{\beta})$ as

$$Q_\lambda(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \\ \sqrt{\lambda}\boldsymbol{\beta}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \\ \sqrt{\lambda}\boldsymbol{\beta}\end{bmatrix}$$

$$= \left\| \begin{bmatrix}\boldsymbol{y} \\ \boldsymbol{0}\end{bmatrix} - \begin{bmatrix}\boldsymbol{X} \\ \sqrt{\lambda}\boldsymbol{I}\end{bmatrix}\boldsymbol{\beta} \right\|_2^2,$$

we observe that ridge regression can be viewed as a standard LS formulation for the augmented problem

$$\begin{bmatrix}\boldsymbol{X} \\ \sqrt{\lambda}\boldsymbol{I}\end{bmatrix}\boldsymbol{\beta} \approx \begin{bmatrix}\boldsymbol{y} \\ \boldsymbol{0}\end{bmatrix}.$$

- The associated normal equations of ridge regression

$$(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y} \tag{6.14}$$

  are obtained from those of original LS problem by adding $\lambda \boldsymbol{I}$ to the coefficient matrix, guaranteeing positive definiteness for $\lambda > 0$.
- Given an SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ of the data matrix $\boldsymbol{X}$ with orthogonal matrices $\boldsymbol{U} = [\boldsymbol{u}_1|\ldots|\boldsymbol{u}_n] \in \mathbb{R}^{n\times n}$, $\boldsymbol{V} = [\boldsymbol{v}_1|\ldots|\boldsymbol{v}_p] \in \mathbb{R}^{p\times p}$ and, and assuming it has full rank $p \le n$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_p \\ \boldsymbol{O} \end{bmatrix}$, $\boldsymbol{\Sigma}_p = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$, $\sigma_1 \ge \cdots \ge \sigma_p > 0$, this implies

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top, \quad \boldsymbol{\Sigma}^\top \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2),$$

$$\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{y} = \sum_{j=1}^{p} \sigma_p (\boldsymbol{u}_j^\top \boldsymbol{y}) \boldsymbol{v}_j$$

.

- Inserting these expressions into the normal equations (6.14) yields

$$\boldsymbol{V}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}+\lambda\boldsymbol{I})\boldsymbol{V}^\top\boldsymbol{\beta}=\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{y}$$

  or, setting $\boldsymbol{\gamma}:=\boldsymbol{V}^\top\boldsymbol{\beta}$,

$$(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}+\lambda\boldsymbol{I})\boldsymbol{\gamma}=\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{y},\qquad\text{giving}\qquad\gamma_j=\frac{\sigma_j}{\sigma_j^2+\lambda}\boldsymbol{u}_j^\top\boldsymbol{y},\quad j=1,\ldots,p,$$

  and finally, with $\boldsymbol{\beta}=\boldsymbol{V}\boldsymbol{\gamma}$, the ridge regression estimate

$$\hat{\boldsymbol{\beta}}_R=\sum_{j=1}^{p}\frac{\sigma_j}{\sigma_j^2+\lambda}(\boldsymbol{u}_j^\top\boldsymbol{y})\boldsymbol{v}_j.$$

- Observe that $\hat{\boldsymbol{\beta}}_R$ is obtained from the standard LS estimate $\hat{\boldsymbol{\beta}}=\sum_{j=1}^{p}\frac{\boldsymbol{u}_j^\top\boldsymbol{y}}{\sigma_j}\boldsymbol{v}_j$ by multiplying each coefficient with the **filter factor**

$$\frac{\sigma_j^2}{\sigma_j^2+\lambda},\qquad j=1,\ldots,p.$$

- Given SVD, ridge regression estimates for additional λ essentially for free.

Principal Components

Covariance matrix of a random vector

- Recall: the variance of a random variable X with expectation $\mu := \mathbf{E}[X]$ is given by

$$\sigma^2 = \mathbf{Var}\, X = \mathbf{E}\left[(X-\mu)^2\right].$$

- For a **random vector** $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ with expectation $\boldsymbol{\mu} := \mathbf{E}[X]$, the **variance** or **covariance matrix** is given by

$$\boldsymbol{C} := \mathbf{Var}\, X = \mathbf{E}\left[(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^\top\right] = \boldsymbol{C}^\top \in \mathbb{R}^{p\times p},$$

with matrix entries

$$C_{i,j} = \mathbf{E}\left[(X_i-\mu_i)(X_j-\mu_j)\right] = \mathbf{Cov}(X_i, X_j), \quad i,j = 1,\ldots,p.$$

Principal Components

Total variance a random vector

- A scalar measure of the total variance contained in a random vector $X \in \mathbb{R}^p$ is provided by the **trace** of its covariance matrix

$$\operatorname{tr} \boldsymbol{C} = \sum_{j=1}^{p} C_{j,j} = \sum_{j=1}^{p} \mathbf{Cov}(X_j, X_j) = \sum_{j=1}^{p} \mathbf{Var}\, X_j.$$

- Justification:

$$\begin{aligned}\mathbf{E}\left[\|X - \mathbf{E}[X]\|_2^2\right] &= \mathbf{E}\left[\|X - \boldsymbol{\mu}\|_2^2\right] = \mathbf{E}\left[(X-\boldsymbol{\mu})^\top (X-\boldsymbol{\mu})\right] \\ &= \mathbf{E}\left[\sum_{j=1}^{p} (X_j - \mu_j)^2\right] = \sum_{j=1}^{p} \mathbf{E}\left[(X_j-\mu_j)^2\right] = \sum_{j=1}^{p} \mathbf{Var}\, X_j.\end{aligned}$$

- By a well-known result from linear algebra, if $\lambda_j(\boldsymbol{C})$ denotes the j-th eigenvalue (in descending order) of $\boldsymbol{C}$[8], there also holds

$$\operatorname{tr} \boldsymbol{C} = \sum_{j=1}^{p} \lambda_j(\boldsymbol{C}).$$

[8] Note that these are real and positive as $\boldsymbol{C}$ is symmetric and positive-definite.

- Given a spectral decomposition

$$\boldsymbol{C} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^\top, \quad \boldsymbol{W}^\top\boldsymbol{W} = \boldsymbol{I}, \quad \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p),$$

of $\boldsymbol{C}$ and the fact that the **Frobenius norm** $\|\cdot\|_F$ is unitarily invariant, we also have

$$\operatorname{tr}\boldsymbol{C} = \|\boldsymbol{\Lambda}^{1/2}\|_F^2 = \|\boldsymbol{W}\boldsymbol{\Lambda}^{1/2}\boldsymbol{W}^\top\|_F^2 = \|\boldsymbol{C}^{1/2}\|_F^2.$$

- In view of the fact that $|\lambda_j(\boldsymbol{C})| = \lambda_j(\boldsymbol{C})$ for covariance matrices, the spectral decomposition $\boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^\top$ is also a singular value decomposition.
- Combining with Theorem 6.3, we conclude that for any $k \in \{1, \ldots, p\}$ the matrix

$$\boldsymbol{C}_k = \sum_{j=1}^{k} \lambda_j \boldsymbol{w}_j \boldsymbol{w}_j^\top,$$

where $\boldsymbol{W} = [\boldsymbol{w}_1 | \ldots | \boldsymbol{w}_p]$, is the best approximation of the covariance matrix $\boldsymbol{C}$ in the spectral and Frobenius norms among all matrices of rank $\leq k$.

Principal Components

Linear combinations of random vector components

- Given a random vector $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ and $\boldsymbol{w}_j$ a normalized eigenvector of its covariance matrix $\boldsymbol{C}$ with associated eigenvalue λ_j, define the scalar random variable $Z_j := \boldsymbol{w}_j^\top X$. Then

$$\begin{aligned}\mathbf{Var}\, Z_j &= \mathbf{E}\left[(\boldsymbol{w}_j^\top X - \mathbf{E}\left[\boldsymbol{w}_j^\top X\right])^2\right] = \mathbf{E}\left[\left(\boldsymbol{w}_j^\top (X-\boldsymbol{\mu})\right)^2\right]. \\ &= \mathbf{E}\left[(\boldsymbol{w}_j^\top (X-\boldsymbol{\mu}))(X-\boldsymbol{\mu})^\top \boldsymbol{w}_j)\right] = \boldsymbol{w}_j^\top \mathbf{E}\left[(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^\top\right] \boldsymbol{w}_j \\ &= \boldsymbol{w}_j^\top \boldsymbol{C}\boldsymbol{w}_j = \lambda_j.\end{aligned}$$

- More generally, for any linear combination $Z = \boldsymbol{\phi}^\top X$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$, we have

$$\begin{aligned}\mathbf{Var}\, Z &= \mathbf{E}\left[(\boldsymbol{\phi}^\top X - \mathbf{E}\left[\boldsymbol{\phi}^\top X\right])^2\right] = \mathbf{E}\left[(\boldsymbol{\phi}^\top (X-\boldsymbol{\mu}))^2\right] \\ &= \mathbf{E}\left[\left(\sum_{j=1}^p \phi_j (X_j - \mu_j)\right)^2\right] = \sum_{j,k=1}^p \phi_j \phi_k \mathbf{E}\left[(X_j-\mu_j)(X_k-\mu_k)\right] \\ &= \boldsymbol{\phi}^\top \boldsymbol{C}\boldsymbol{\phi}.\end{aligned}$$

Principal Components

Linear combinations of random vector components

- For two general linear combinations $Z_1 = \boldsymbol{\phi}_1^\top X,\ Z_2 = \boldsymbol{\phi}_2^\top X$, we conclude by an analogous calculation that

$$\mathbf{Cov}(Z_1, Z_2) = \boldsymbol{\phi}_2^\top \boldsymbol{C} \boldsymbol{\phi}_1$$

  and therefore that Z_1 and Z_2 are **uncorrelated** if and only if $\boldsymbol{\phi}_2^\top \boldsymbol{C} \boldsymbol{\phi}_1 = 0$, i.e., if the coefficient vectors $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are orthogonal in the inner product generated by the (symmetric and positive definite) matrix $\boldsymbol{C}$.
- If we seek a change of variables $Z = \boldsymbol{\Phi}^\top X$ with a nonsingular $\boldsymbol{\Phi} \in \mathbb{R}^{p\times p}$ such that the components of Z are uncorrelated with unit variance, then it is necessary that

$$\boldsymbol{I} = \mathbf{E}\left[(Z - \mathbf{E}[Z])(Z - \mathbf{E}[Z])^\top\right] = \mathbf{E}\left[\boldsymbol{\Phi}^\top (X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^\top \boldsymbol{\Phi}\right] = \boldsymbol{\Phi}^\top \boldsymbol{C} \boldsymbol{\Phi}.$$

  The set of all matrices $\boldsymbol{\Phi} \in \mathbb{R}^{p\times p}$ which achieve this is precisely the set of all **congruences** taking $\boldsymbol{C}$ to $\boldsymbol{I}$.

Principal Components

Linear combinations of random vector components

- **Example 1:** given **Cholesky factorization** $C = LL^\top$, choosing $\Phi := L^{-\top}$ gives
$$\Phi^\top C \Phi = L^{-1}(LL^\top)L^{-\top} = I.$$
- **Example 2:** given **spectral decomposition** $C = W\Lambda W^\top$, choosing $\Phi := W\Lambda^{-1/2}$ gives
$$\Phi^\top C \Phi = \Lambda^{-1/2} W^\top (W\Lambda W^\top) W\Lambda^{-1/2} = I.$$
- **Example 3:** given **square-root-free Cholesky factorization** $C = LDL^\top$, where L is lower triangular with a unit diagonal and D is diagonal, choosing $\Phi := L^{-\top}$ gives
$$\Phi^\top C \Phi = L^{-1}(LDL^\top)L^{-\top} = D.$$
- **Example 4:** given **spectral decomposition** $C = W\Lambda W^\top$, choosing $\Phi := W$ gives
$$\Phi^\top C \Phi = W^\top (W\Lambda W^\top) W = \Lambda.$$

For a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the expression $\dfrac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}$, $\quad \boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^n$, is called a **Rayleigh quotient**.

Theorem 6.4 (Fischer, 1905; Courant, 1920)

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $k \in \{1, 2, \ldots, n\}$. Then

$$\lambda_k = \min_{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{n-k} \in \mathbb{R}^n} \; \max_{\substack{0\neq \boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{x}\perp \boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{n-k}}} \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}, \tag{6.15}$$

$$\lambda_k = \max_{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{k-1} \in \mathbb{R}^n} \; \min_{\substack{0\neq \boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{x}\perp \boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{k-1}}} \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \tag{6.16}$$

- The extremal values of the Rayleigh quotient are attained when $\boldsymbol{x}$ is an eigenvector associated with λ_1 or λ_n, respectively.

Consequences of Theorem 6.4:

- Linear combination $\boldsymbol{\phi}^\top X$ where $\|\boldsymbol{\phi}\|_2 = 1$ with maximal variance obtained for $\boldsymbol{\phi} = \boldsymbol{\phi}_1 = \boldsymbol{w}_1$. This is the first principal component.
- Linear combination $\boldsymbol{\phi}^\top X$ where $\|\boldsymbol{\phi}\|_2 = 1$ with maximal variance subject to $\boldsymbol{\phi} \perp \boldsymbol{w}_1$ obtained for $\boldsymbol{\phi} = \boldsymbol{\phi}_2 = \boldsymbol{w}_2$ (second principal component).
- Linear combination $\boldsymbol{\phi}^\top X$ where $\|\boldsymbol{\phi}\|_2 = 1$ with maximal variance subject to $\boldsymbol{\phi} \perp \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{j-1}$ obtained for $\boldsymbol{\phi} = \boldsymbol{\phi}_j = \boldsymbol{w}_j$ (j-th principal component).
- The change of variables afforded by replacing the original random variables $X_1, \ldots, X_p$ by the principal components $Z = \boldsymbol{W}^\top X$ is the (unscaled) congruence obtained from the spectral decomposition.
  The total variance contained in Z is given by

$$\mathbf{E}\left[\|Z - \mathbf{E}[Z]\|_2^2\right] = \sum_{j=1}^{p} \mathbf{Var}\, Z_j = \sum_{j=1}^{p} \lambda_j = \operatorname{tr} \sum_{j=1}^{p} \lambda_j \boldsymbol{w}_j \boldsymbol{w}_j^\top = \operatorname{tr} \boldsymbol{C},$$

  which coincides with the total variance contained in X.

Principal Components

PCR

- Performing regression of a data vector $\boldsymbol{y}$ on $M < p$ principal components results in **principal components regression** (PCR).
- The total variance contained in random vector $(Z_1, \ldots, Z_M)^\top$ is

$$\mathbf{E}\left[\|Z - \mathbf{E}[Z]\|_2^2\right] = \sum_{j=1}^{M} \mathbf{Var}\, Z_j = \sum_{j=1}^{M} \lambda_j = \operatorname{tr} \sum_{j=1}^{M} \lambda_j \boldsymbol{w}_j \boldsymbol{w}_j^\top = \operatorname{tr} \boldsymbol{C}_k.$$

- The fraction of neglected variance in PCR using M principal components is

$$\frac{\sum_{j=k+1}^{p} \lambda_j}{\sum_{j=1}^{p} \lambda_j} = 1 - \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

- The covariance matrix $\boldsymbol{C}$ and expectation vector $\boldsymbol{\mu}$ are theoretical constructs and typically unavailable hence estimated from data.
- As usual, we denote the data matrix (design matrix) by

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix} = [\boldsymbol{x}_1 | \cdots | \boldsymbol{x}_p] \in \mathbb{R}^{n \times p},$$

  each column corresponding to one of p predictor variables (features) and each row to one of n observations (samples, realizations).
- We denote the vector of **sample means** by $\overline{\boldsymbol{X}} := \frac{1}{n}\boldsymbol{e}^\top \boldsymbol{X} = [\overline{x_1}, \ldots, \overline{x_p}]$ and obtain the **centered data matrix** as

$$\tilde{\boldsymbol{X}} := [\boldsymbol{x}_1 - \overline{x_1}\boldsymbol{e} | \cdots | \boldsymbol{x}_p - \overline{x_p}\boldsymbol{e}] = \boldsymbol{X} - \boldsymbol{e}\overline{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{e}\frac{1}{n}\boldsymbol{e}^\top \boldsymbol{X} = (\boldsymbol{I} - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^\top)\boldsymbol{X}.$$

- Finally, the **unbiased sample covariance matrix** is

$$\boldsymbol{S}_n := \frac{1}{n-1}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} = \frac{1}{n-1}\boldsymbol{X}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^\top)^2\boldsymbol{X} = \frac{1}{n-1}\boldsymbol{X}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^\top)\boldsymbol{X}.$$

Principal Components

Data

- In practice the sample covariance matrix $\boldsymbol{S}_n$ takes the place of the covariance matrix $\boldsymbol{C}$.
- For PCA/PCR, one can compute a spectral decompositon of $\boldsymbol{S}_n$.
- Alternatively, given an SVD $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, a spectral decomposition of $\boldsymbol{S}_n$ is obtained as

$$\boldsymbol{S}_n = \frac{1}{n-1}\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} = \frac{1}{n-1}\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

- The SVD approach is generally numerically stabler, in particular if $\tilde{\boldsymbol{X}}$ is ill-conditioned. The spectral decomposition may be cheaper, as $\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}$ is smaller than $\tilde{\boldsymbol{X}}$.