Unit 2

Regularized regression: lasso, ridge and elastic net

Linear regression models and least squares

- The linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j,$$

  with input vector $X^T = (X_1, \ldots, X_p)$.
- The linear model assumes that $\mathrm{E}(Y|X)$ is linear or that the linear model is a reasonable approximation.
- The variables X_j can come from different sources:
  - quantitative inputs;
  - transformations of quantitative inputs;
  - basis expansions;
  - numeric or “dummy” coding of levels of qualitative inputs;
  - interactions between variables.

Linear regression models and least squares / 2

- The model is linear in the (unknown) parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$.
- Least squares estimation determines β by minimizing the residual sum of squares:

$$\begin{aligned} \text{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_i))^2 \\ &= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \end{aligned}$$

  In matrix notation:

$$\text{RSS}(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta).$$

Linear regression models and least squares / 3

- Differentiating with respect to β gives:

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta),$$
$$\frac{\partial^2 \text{RSS}}{\partial \beta \partial \beta^T} = 2\boldsymbol{X}^T\boldsymbol{X}.$$

- Assuming that $\boldsymbol{X}$ has full column rank, $\boldsymbol{X}^T\boldsymbol{X}$ is positive definite and the unique solution is given by

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

Linear regression models and least squares / 4

- The predicted values are given by

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

  with $\boldsymbol{H}$ the "hat"-matrix which orthogonally projects into the subspace spanned by the columns of $\boldsymbol{X}$.
- If $\boldsymbol{X}^T\boldsymbol{X}$ is singular, β is not uniquely determined, but the fitted values $\hat{\boldsymbol{y}}$ are.

Linear regression models and least squares / 5

- Assume

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon,$$

with

1. $\mathrm{E}(\epsilon) = \mathbf{0}$
2. $\mathrm{Var}(\epsilon) = \sigma^2 \boldsymbol{I}$
3. $\boldsymbol{X}$ deterministic with full column rank.

- Then:
  - The Gauss-Markov Theorem gives that the least squares estimator $\hat{\beta}$ is the best linear unbiased estimator (BLUE).
  - The variance of the estimator is:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Linear regression models and least squares / 6

- Typically the variance σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,$$

  to obtain an unbiased estimate.

- Adding the assumption:

  4. $\epsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$

  gives

$$\hat{\beta} \sim N(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}),$$
$$(N-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p-1},$$

  and $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent.

Linear regression models and least squares / 7

- The t test statistic for the hypothesis test $\beta_j = 0$ is given by

$$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}},$$

with v_j the jth diagonal element of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Under the Gaussian assumption and the null hypothesis $\beta_j = 0$ t_j follows a t distribution with $N - p - 1$ degrees of freedom.

- This can also be used to construct $(1 - 2\alpha)$ confidence intervals for β_j by:

$$(\hat{\beta}_j - t_{N-p-1}^{(1-\alpha)}\sqrt{v_j}\hat{\sigma}, \hat{\beta}_j + t_{N-p-1}^{(1-\alpha)}\sqrt{v_j}\hat{\sigma}),$$

with $t_{N-p-1}^{(1-\alpha)}$ the $1 - \alpha$ quantile of the t distribution with $N - p - 1$ degrees of freedom.

Linear regression models and least squares / 8

- To compare two nested models, one calculates

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p_1 - p_0)}{\text{RSS}_1/(N - p_1 - 1)},$$

with

  - RSS_1 being the residual sum-of-squares for the least squares fit of the bigger model with $p_1 + 1$ parameters, and
  - RSS_0 the same for the nested smaller model with $p_0 + 1$ parameters, having $p_1 - p_0$ parameters constrained to be zero.

Under the Gaussian assumption and that the smaller model is correct, the F statistic follows an F-distribution with $p_1 - p_0$ and $N - p_1 - 1$ degrees of freedom.

Estimation in R

The function for fitting a linear regression model in R is:

- The formula specifies the dependent variable as well as how the model matrix is created from the independent variables.
- Least squares estimation is performed using QR method.
- Regression coefficient estimates are identical to maximum likelihood estimates assuming that ϵ is normally distributed:

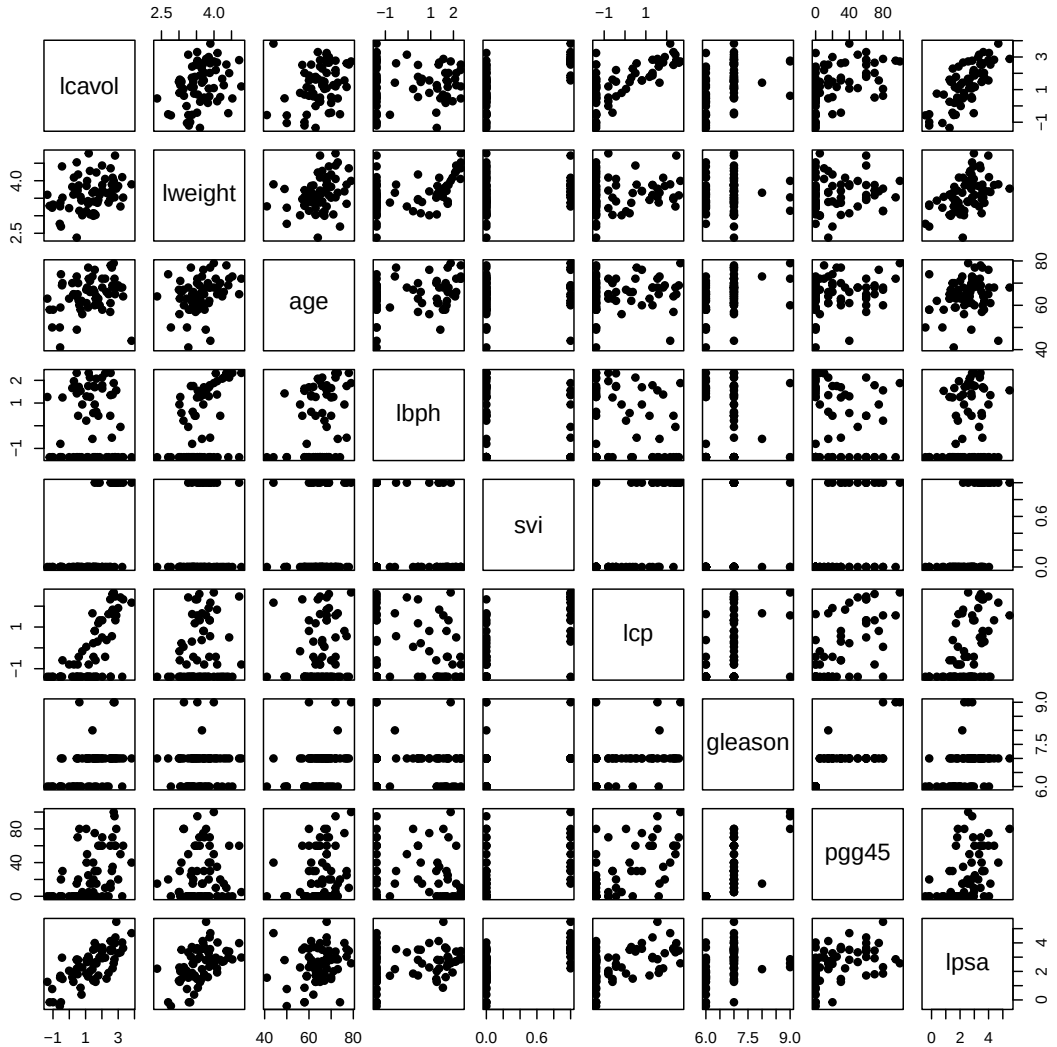
$$\epsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

  Inference performed (e.g., using `summary()`) is based on this assumption.

Example: Prostate Cancer

- Data set provided in the R package **ElemStatLearn**.
- Use only the 67 observations from the training data set.
- The dependent variable is `lpsa`, the level of a prostate-specific antigen.
- The independent variables are clinical measures.
- There is a substantial amount of correlation between the clinical measures.
- The independent variables are standardized before fitting the linear model.

Example: Prostate Cancer / 2



Example: Prostate Cancer / 3

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.45       0.09   28.18   <2e-16 ***
lcavol          0.72       0.13    5.37   <2e-16 ***
lweight         0.29       0.11    2.75     0.01 **
age            -0.14       0.10   -1.40     0.17
lbph            0.21       0.10    2.06     0.04 *
svi             0.31       0.13    2.47     0.02 *
lcp            -0.29       0.15   -1.87     0.07 .
gleason        -0.02       0.14   -0.15     0.88
pgg45           0.28       0.16    1.74     0.09 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

Example: Prostate Cancer / 4

- Dropping the insignificant variables (`age`, `lcp`, `gleason`, `pgg45`) and comparing the two models using the F-test gives:

```
Analysis of Variance Table

Model 1: lpsa ~ (lcavol + lweight + age + lbph + svi + lcp +
    pgg45) - age - lcp - gleason - pgg45
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    pgg45
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     62 32.815
2     58 29.426  4    3.3886 1.6698 0.1693
```

- Prediction performance on the test dataset measured by the MSE:

```
base model   full model subset model
     1.057        0.521        0.456
```

Model / subset selection

- There are two reasons why least squares estimates are not satisfactory:
  - *Prediction accuracy*: Shrinking or setting coefficients zero might induce bias, but may improve prediction accuracy due to reduced variance.
  - *Interpretation*: Determining a small subset of predictors exhibiting the strongest effects allows to more easily discern important regressors.

Model / subset selection / 2

Strategies for choosing a subset:

- Best subset selection: possible for up to $p = 30$ or 40 using the efficient *leaps and bounds* procedure (Furnival and Wilson, 1974). Available for example in package **leaps** in R.
- Forward- and backward-stepwise selection: seek a good path through subsets of different size.
  - Forward selection starts with the intercept model and iteratively adds predictors which improve the model fit most.
  - Backward selection starts with the full model and iteratively removes the predictor which has the least impact on the fit.
  - Variables are iteratively added or removed. In each step the least squares coefficient estimates are determined for the current *active set*.

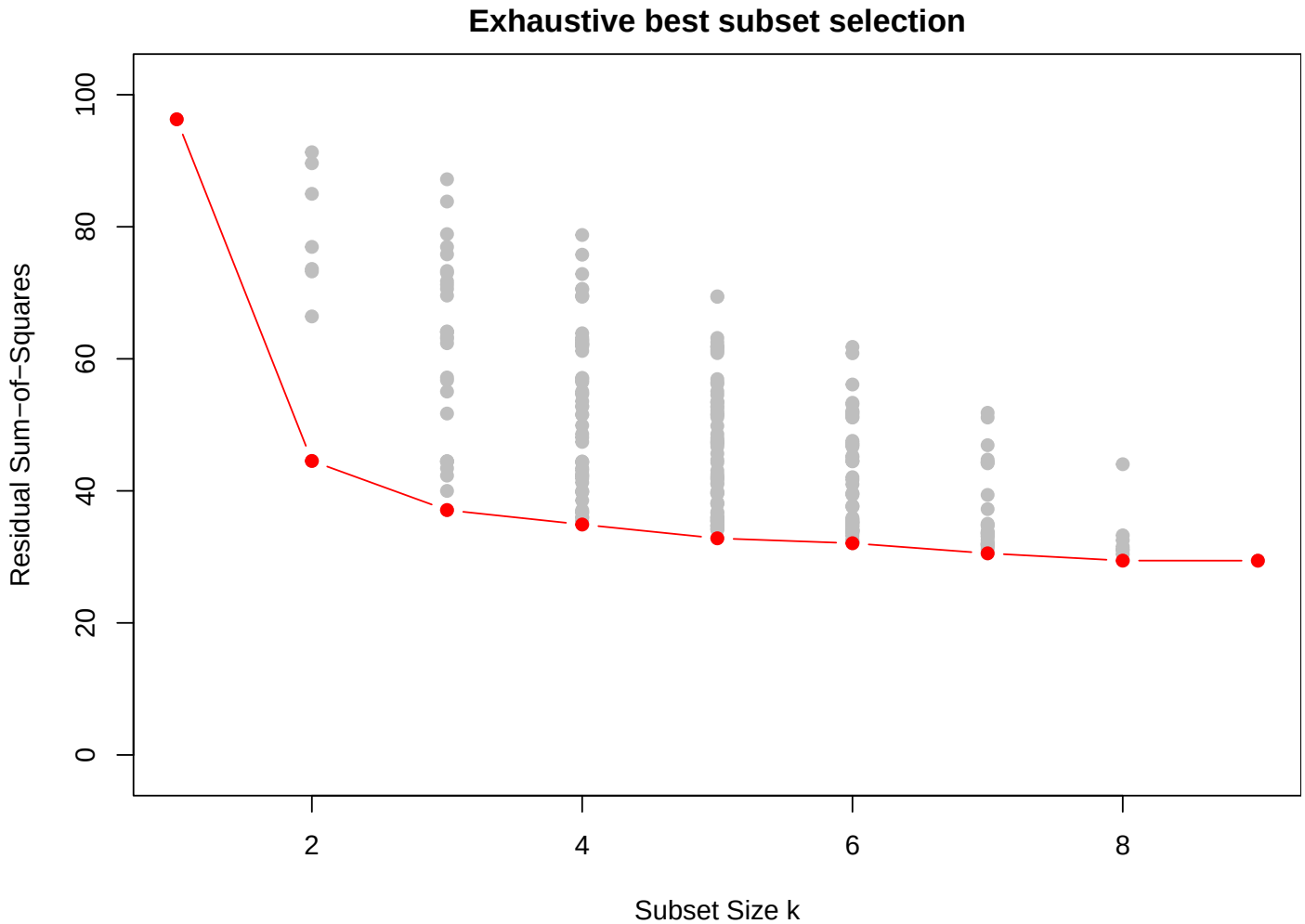
Model / subset selection / 3

- Forward-stagewise regression:
  - In each step the variable is selected which has the largest absolute correlation with the current residual.
  - The simple linear regression coefficient for this variable and the current residual is determined.
  - The coefficients of all other variables remain unchanged.
  - More than p steps necessary to arrive at the full least squares solution.

Model / subset selection / 4

- Least angle regression (LAR):
  - Iterative procedure where one variable is added in each step to the *active set*.
  - Starting at each step the variables in the active set
    - have the same absolute correlation with the current residual,
    - have a higher absolute correlation than the variables not in the active set.
  - The coefficients of the variables in the active set are updated
    - keeping the correlation to the residuals tied,
    - until another variable (not in the active set) has the same correlation with the residuals.
  - This variable is then added to the active set.
  - This process eventually also gives the full least squares solution.

Example: Prostate Cancer



Shrinkage methods

- Shrinkage methods add a complexity parameter which allows to gradually change between a simple model (e.g., intercept only) to a complex model (e.g., the least squares fit of all variables).
- In general the optimization criterion is modified by adding a penalty.
- Examples:
  - Best subset selection
  - Ridge
  - Lasso (least absolute shrinkage and selection operator)
  - Elastic net

Best subset selection

- Regression with best subset selection obtains the estimates by adding a penalty in dependence of the L_0 norm of the regression coefficients:

$$\hat{\beta}^{\text{BSS}} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \|\beta_{\setminus 0}\|_0 \right\},$$

  with the L_0 norm being equal to the number of non-zero elements in a vector and with $\lambda \geq 0$ the complexity parameter.
- An equivalent problem formulation is:

$$\hat{\beta}^{\text{BSS}} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\},$$

$$\text{subject to } \|\beta_{\setminus 0}\|_0 \leq t,$$

  with a one-to-one correspondence between t and λ.

Ridge

- Ridge regression obtains the estimates by adding a penalty in dependence of the squared Euclidean length of the regression coefficients:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

  with $\lambda \geq 0$ the complexity parameter:
  - $\lambda = 0$: least squares fit.
  - $\lambda = \infty$: only a constant function β_0 fit.

Ridge / 2

- An equivalent problem formulation is:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\},$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t,$$

  with a one-to-one correspondence between t and λ.
- The intercept β_0 is not penalized to ensure that the solutions do not depend on the origin for Y.
- Ridge regression alleviates the problem of poorly identified coefficients in case of multicollinearity.
- Ridge regression is not equivariant under scaling of the inputs.
  $\Rightarrow$ Normally inputs are standardized before estimation.

Ridge / 3

- The ridge criterion can also be written as

$$\text{RSS}(\beta) = \frac{1}{N}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\beta^T\beta,$$

with solution

$$\hat{\beta}^{\text{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda N\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix.

Ridge: Shrinkage effects

- For orthonormal inputs where $\boldsymbol{X}^T\boldsymbol{X} = N\boldsymbol{I}$:

$$\hat{\beta}^{\text{ridge}} = \frac{\hat{\beta}}{1+\lambda}.$$

- Otherwise use the singular value decomposition

$$\frac{1}{\sqrt{N}}\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T,$$

with

  - $\boldsymbol{U}$ a $N \times p$ orthogonal matrix,
  - $\boldsymbol{V}$ a $p \times p$ orthogonal matrix,
  - $\boldsymbol{D}$ a $p \times p$ diagonal matrix with $d_1 \geq d_2 \geq \ldots d_p \geq 0$.

Ridge: Shrinkage effects / 2

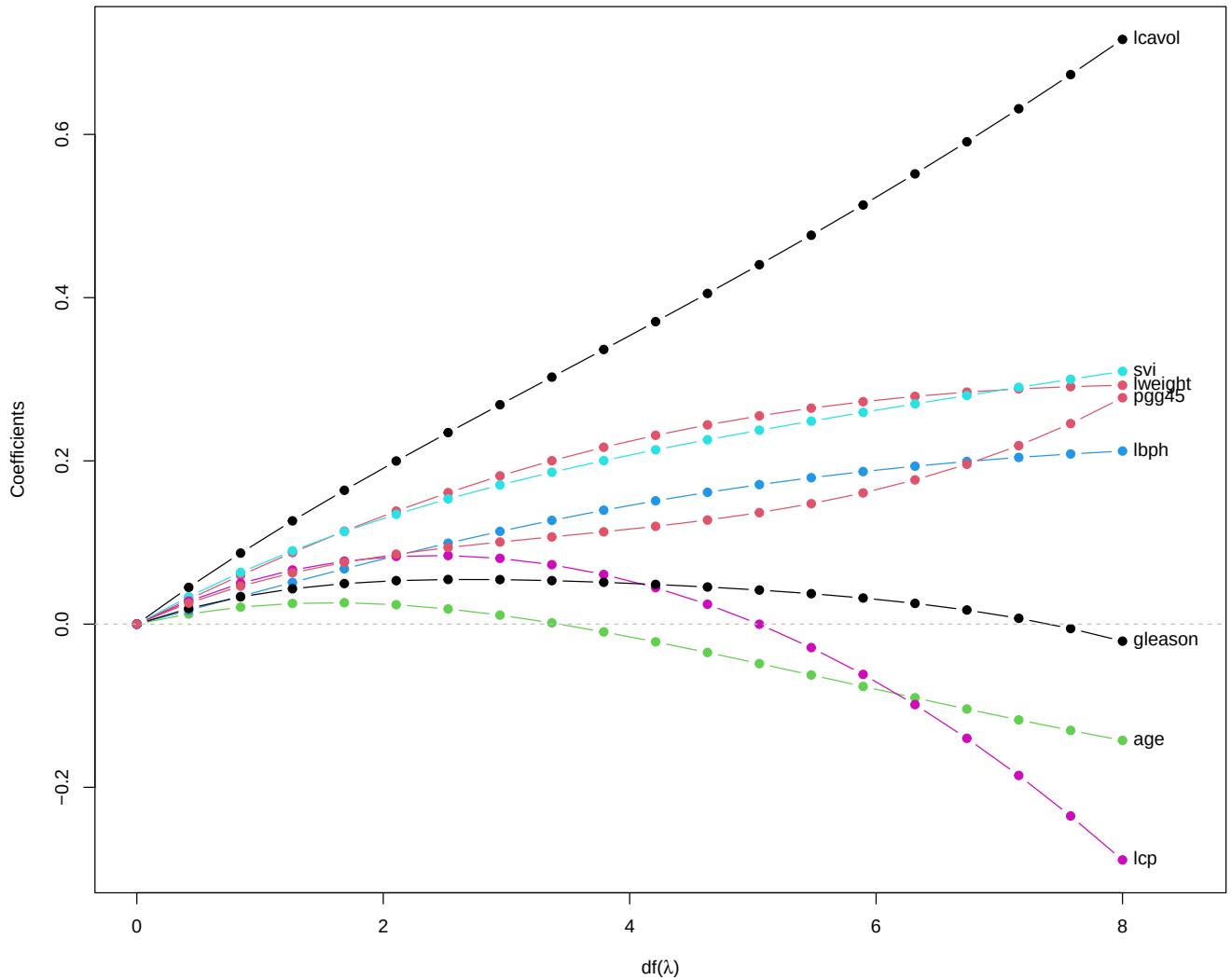
- This gives

$$\begin{aligned} \boldsymbol{X}\hat{\beta}^{\text{ridge}} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda N\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y}. \end{aligned}$$

- Ridge regression
  - projects $\boldsymbol{y}$ onto the principal components and
  - shrinks the coefficients of the low-variance components more than the high-variance components.
- The *effective degrees of freedom* are given by

$$\begin{aligned} \text{df}(\lambda) &= \text{tr}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda N\boldsymbol{I})^{-1}\boldsymbol{X}^T] = \text{tr}(\boldsymbol{H}_\lambda) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}. \end{aligned}$$

Example: Prostate Cancer



Lasso

- The lasso estimate is defined by:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\},$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

- The equivalent *Lagrangian form* is given by:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

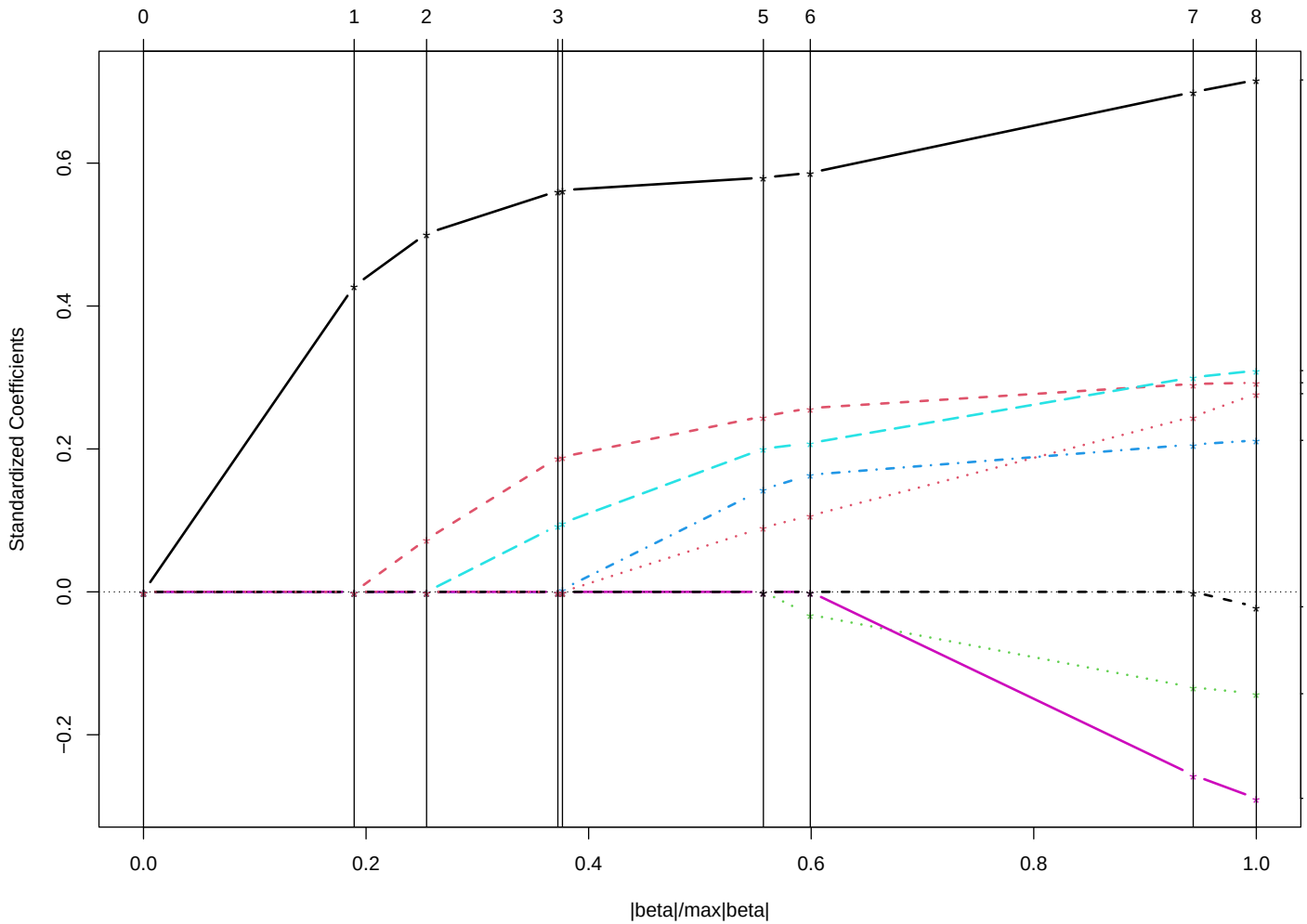
Lasso / 2

- The lasso essentially replaces the L_2 ridge penalty with the L_1 penalty.
- If t is sufficiently small, some of the coefficients will be exactly zero.
- If t is larger than $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$, then the lasso estimates are the OLS estimates $\hat{\beta}_j$. If $t = t_0/2$, the lasso estimates are shrunken by 50% on average.
- A standardized penalty parameter is given by

$$s = \frac{t}{\sum_{j=1}^{p} |\hat{\beta}_j|},$$

  with $s \in [0, 1]$.
- Lasso regression is not equivariant under scaling of the inputs.
  $\Rightarrow$ Normally inputs are standardized before estimation.

Example: Prostate Cancer



Elastic net

The following optimization problem is solved in elastic net regression

$$\arg\min_{\beta} \left\{ \frac{1}{2N} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \left(\frac{1}{2}(1-\alpha)\|\beta_{[-0]}\|_2^2 + \alpha\|\beta_{[-0]}\|_1 \right) \right\} =$$

$$\arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\sum_{j=1}^{p} \frac{1-\alpha}{2}\beta_j^2 + \alpha|\beta_j| \right) \right\}.$$

- The parameter $\lambda \geq 0$ is a complexity parameter, that controls the amount of shrinkage. The parameter $\alpha \in [0, 1]$ determines the compromise between ridge and lasso penalty.
- The intercept β_0 is not shrunken.
- The elastic net solutions are not equivariant under scaling of the regressors.
  $\Rightarrow$ One normally standardizes the regressors before analysis.

Comparison

For an input matrix $\boldsymbol{X}$ where $\boldsymbol{X}^T\boldsymbol{X} = N\boldsymbol{I}$:

- Best subset (size M), L_0 loss:

$$\hat{\beta}_j \mathbb{1}_{\text{rank}(|\hat{\beta}_j|) \leq M}$$

- Lasso, L_1 loss:

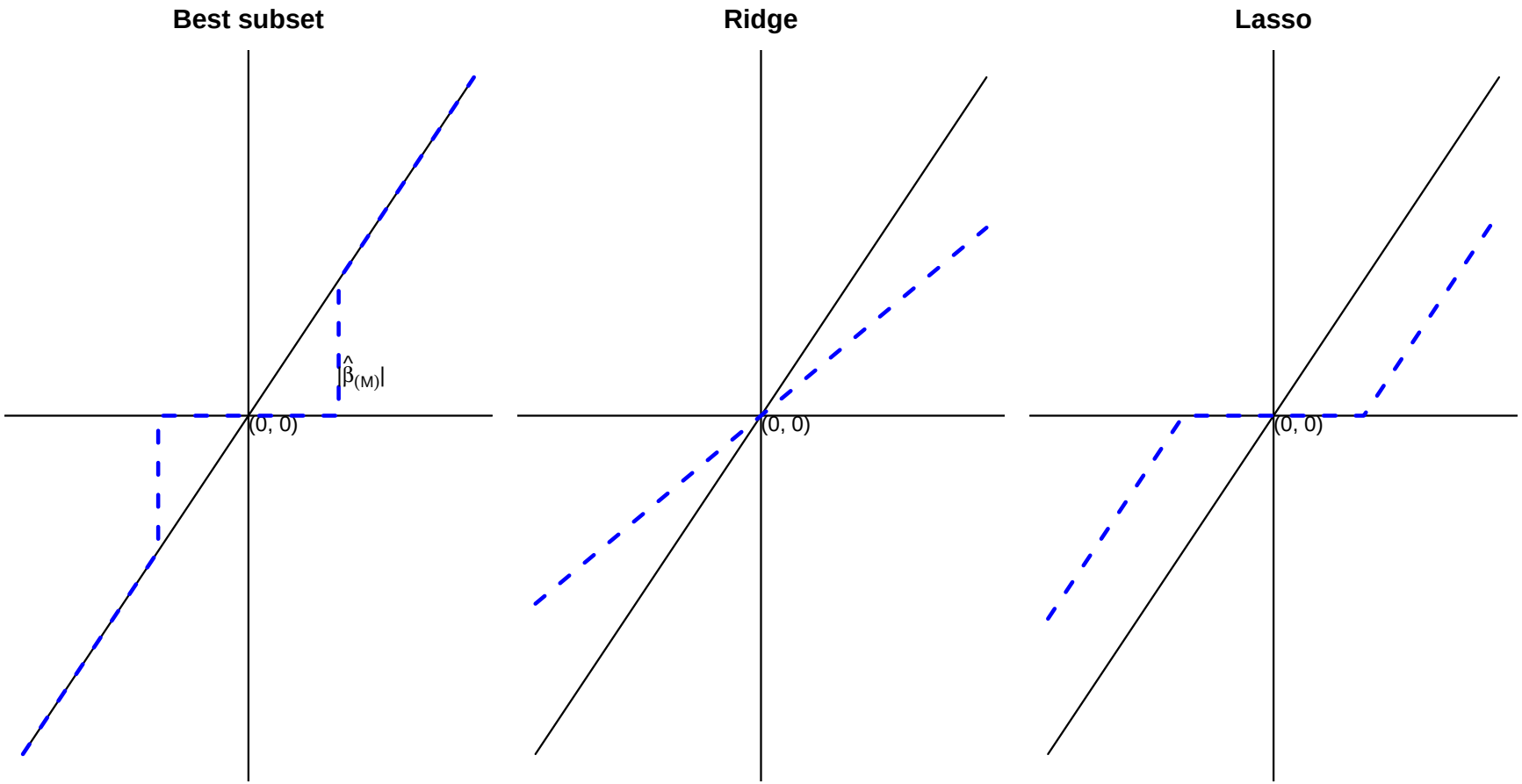
$$\text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$$

- Ridge, L_2 loss:

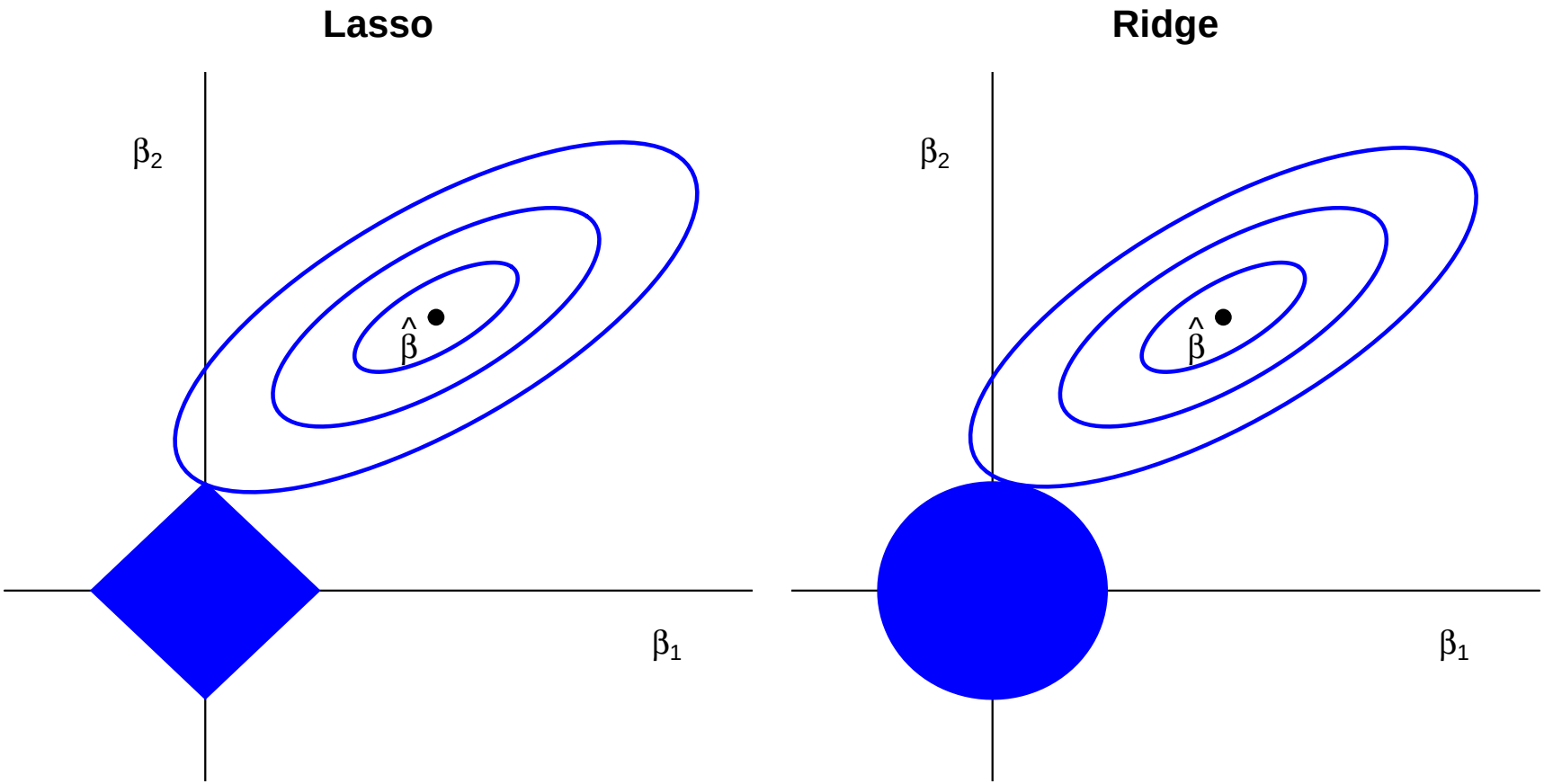
$$\hat{\beta}_j \frac{1}{1+\lambda}$$

with $\hat{\beta}_j$ the OLS estimator.

Comparison / 2



Comparison / 3



Comparison / 4

- The lasso and ridge regression can be generalized to:

$$\tilde{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\},$$

with $q \geq 1$.

Comparison / 5

- Elastic net:

$$\hat{\beta}^{\text{elastic net}} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left(\frac{1-\alpha}{2} \beta_j^2 + \alpha |\beta_j| \right) \right\},$$

with $\alpha \in [0, 1]$.

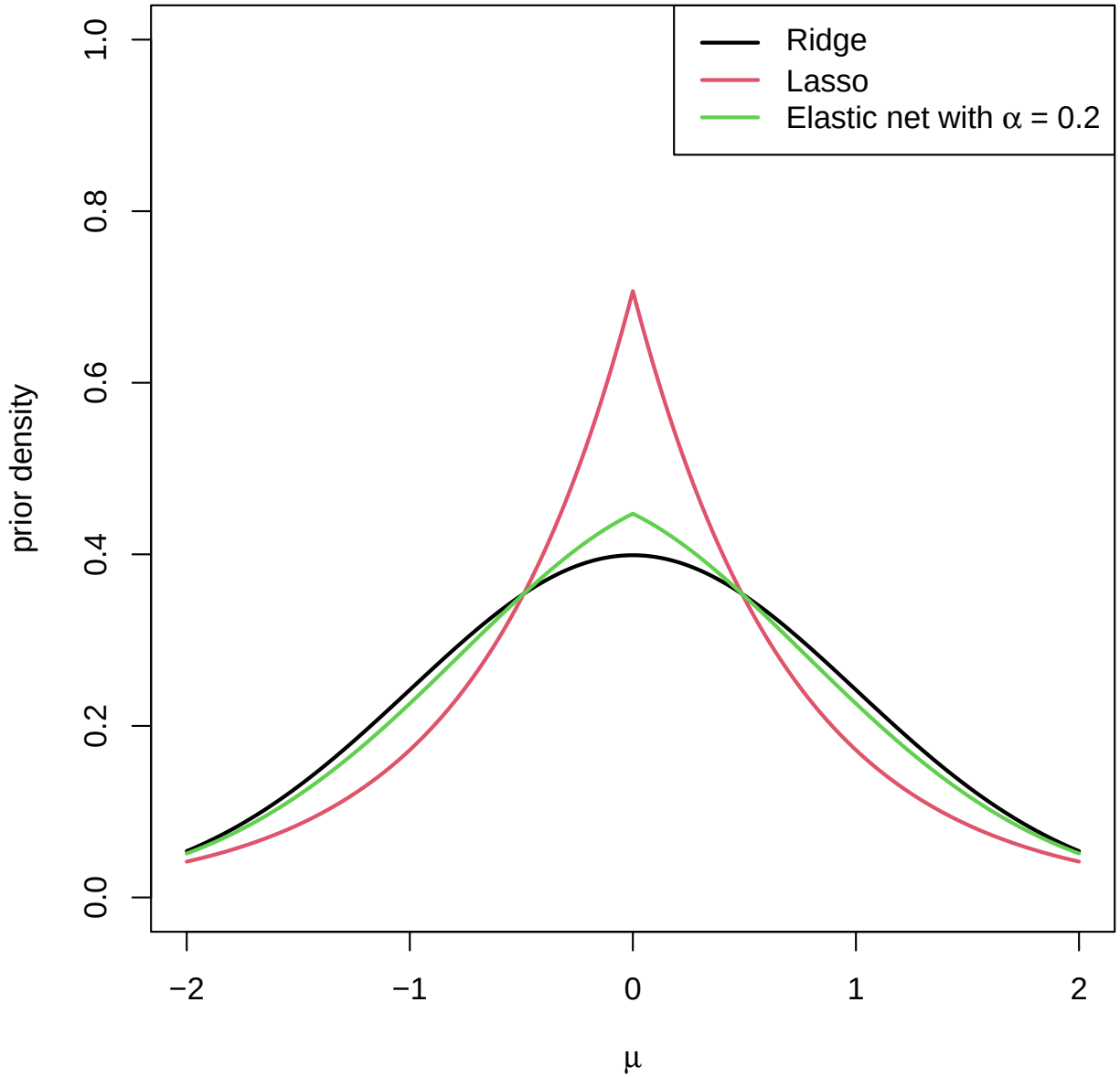
  - Represents a compromise between lasso and ridge.
  - Selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge.
  - Has computational advantages over the L_q penalties.

Relation to Bayesian estimation

- Bayesian estimation determines the posterior distribution of a parameter given
  - prior beliefs and
  - observed data

  by combining the prior distribution of the parameter with the likelihood function.
- All penalized approaches can be seen as determining the maximum a-posteriori estimates of the parameters using different prior distributions. E.g., the following priors for:
  - Ridge: normal distribution.
  - Lasso: double-exponential or Laplace distribution.

Relation to Bayesian estimation / 2



Shrinkage methods: Estimation

- Best subset selection:
  - Use of an efficient branch and bound algorithm to avoid enumeration of all subsets.
  - Exploits that in linear regression it holds for the residual sum of squares (RSS) that

$$\text{RSS}(A) \leq \text{RSS}(B),$$

    where A is any set of independent variables and B is a subset of A.

Shrinkage methods: Estimation / 2

- Forward stagewise regression:
  - Transform the input vectors such that they have mean zero and unit norm, i.e.,

$$\sum_{i=1}^{N} z_{ij} = 0, \qquad \sum_{i=1}^{N} z_{ij}^2 = 1.$$

    Keep the squared length $\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2$ to rescale the final coefficients.
  - Transform the output to have mean zero.

Shrinkage methods: Estimation / 3

- The input vector with the highest correlation to the current residual is the vector where the absolute value of the inner product between transformed input vector and the current residual is maximum.
- Update the coefficient by adding the inner product.
- Update the residual by subtracting the inner product times the transformed input.
- Stop when the absolute value of the inner product is smaller than ϵ.
- Rescale the coefficients using the squared length.

Shrinkage methods: Estimation / 4

- Ridge:
  - Regression coefficient estimates are available in closed form for a given λ.
- Lasso:
  - If λ decreases the coefficient values change in a piecewise linear fashion. The slope only changes if coefficients leave or enter the set of active coefficients.
  - The entire path for all λ values can be determined in a computationally efficient way.
    $\Rightarrow$ See least angle regression (LAR).
  - Pathwise coordinate optimization.
  - Convex optimization problem (as ridge regression).

Least angle regression (LAR)

- Related to:
  - Forward-stepwise regression.
  - Forward-stagewise regression.
  - Lasso.
- Iterative procedure where one variable is added in each step to the *active set*.

Least angle regression (LAR) / 2

- Starting at each step the variables in the active set
  - have the same absolute correlation with the current residual,
  - have a higher absolute correlation than the variables not in the active set.
- The regression coefficients of the variables in the active set are updated
  - keeping the correlation to the residuals tied,
  - until another variable (not in the active set) has the same correlation with the residuals.
- This variable is added to the active set and its regression coefficient also varied.
- This process eventually also gives the full least squares solution.

Least angle regression (LAR) / 3

- Suppose $\mathcal{A}_k$ is the active set of variables at the beginning of the kth step.
- Let $\beta_{\mathcal{A}_k}$ be the coefficient vector for these variables at this step.
- There will be $k - 1$ non-zero values and the one just entered will be zero.
- If $\boldsymbol{r}_k = \boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$ is the current residual, then the direction for this step is:

$$\delta_k = (\boldsymbol{X}_{\mathcal{A}_k}^T \boldsymbol{X}_{\mathcal{A}_k})^{-1} \boldsymbol{X}_{\mathcal{A}_k}^T \boldsymbol{r}_k.$$

- The coefficient profile then evolves as

$$\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha\delta_k.$$

Least angle regression (LAR) / 4

- The exact step length α is given by

$$\min_{j \in \mathcal{A}_k^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{\hat{A} - \hat{a}_j}, \frac{\hat{C} + \hat{c}_j}{\hat{A} + \hat{a}_j}, \right\},$$

  where $\min^+$ indicates that the minimum is taken over only positive components within each choice of j and

$$\begin{aligned}
\hat{\boldsymbol{c}} &= \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_k}\beta_{\mathcal{A}_k}) = \boldsymbol{X}^T\boldsymbol{r}_k, \\
\hat{C} &= \max_{j \in \mathcal{A}_k}\{|\hat{\boldsymbol{c}}_j|\}, \\
\hat{\boldsymbol{a}} &= \boldsymbol{X}^T\boldsymbol{X}_{\mathcal{A}_k}\delta_k, \\
\hat{A} &= \max_{j \in \mathcal{A}_k}\{|\hat{\boldsymbol{a}}_j|\}.
\end{aligned}$$

LAR and lasso

- Assume the input features and the output are standardized.
- For the active set $\mathcal{A}_k$ at step k in LAR it holds:

$$\frac{1}{N}\boldsymbol{x}_j^T(\boldsymbol{y} - \boldsymbol{X}\beta) = \gamma \cdot s_j,$$

for all $j \in \mathcal{A}_k$ and with

- β the current coefficient estimates,
- $s_j \in \{-1, 1\}$,
- γ the common value.

LAR and lasso / 2

- The lasso criterion corresponds to

$$R(\beta) = \frac{1}{2N}\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1.$$

- If $\mathcal{B}$ is the active set of variables for a given value of λ, $R(\beta)$ is differentiable for these variables and the stationarity conditions need to hold:

$$\frac{1}{N}\boldsymbol{x}_j^T(\boldsymbol{y} - \boldsymbol{X}\beta) = \lambda \cdot \text{sign}(\beta_j),$$

  for all $j \in \mathcal{B}$.

LAR and lasso / 3

- The LAR and lasso criteria are identical if the sign of β_j matches the sign of the inner product.
  $\Rightarrow$ LAR and lasso start to differ when an active coefficient passes through zero.
  $\Rightarrow$ For the lasso the variable is excluded from the active set.
- For the non-active variables the stationarity conditions require

$$|\frac{1}{N}\boldsymbol{x}_k^T(\boldsymbol{y} - \boldsymbol{X}\beta)| \leq \gamma$$

for $k \notin \mathcal{A}$ and

$$|\frac{1}{N}\boldsymbol{x}_k^T(\boldsymbol{y} - \boldsymbol{X}\beta)| \leq \lambda$$

for $k \notin \mathcal{B}$.

Algorithm: Least angle regression

1. Transform the predictors to have mean zero and unit norm. Start with the residual $\boldsymbol{r} = \boldsymbol{y} - \bar{y}\mathbf{1}$ and $\beta_1, \ldots, \beta_p = 0$.
2. Find the predictor $\boldsymbol{x}_j$ most correlated with $\boldsymbol{r}$.
3. Move β_j from 0 towards its least squares (LS) coefficient $\boldsymbol{x}_j^T \boldsymbol{r}$, until some other competitor $\boldsymbol{x}_k$ has as much correlation with the current residual as does $\boldsymbol{x}_j$.
4. Move β_j and β_k in the direction defined by their joint LS coefficient of the current residual on $(\boldsymbol{x}_j, \boldsymbol{x}_k)$, until some other competitor $\boldsymbol{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After $\min(N-1, p)$ steps, we arrive at the full LS solution.

Lasso modification:

4a. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint LS direction.

Pathwise coordinate optimization

- An alternate approach to the lars algorithm for computing the lasso solution.
- Fix the penalty parameter λ in the Lagrangian form

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- Optimize successively over each parameter, holding the other parameters fixed at their current values.
- This algorithm can be used to efficiently calculate the lasso solutions at a grid of values of λ. One starts with the largest value of λ.
- This algorithm can also be modified to be used with the elastic net.

Pathwise coordinate optimization: Algorithm

- Suppose all predictors are standardized to have mean zero and unit norm and the response also has mean zero.
- The current estimate for β_k at penalty parameter λ is denoted by $\tilde{\beta}_k(\lambda)$.
- The function to optimize can be written as:

$$R(\tilde{\beta}(\lambda), \beta_j) = \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{k \neq j} x_{ik} \tilde{\beta}_k(\lambda) - x_{ij} \beta_j \right)^2 + \lambda \sum_{k \neq j} |\tilde{\beta}_k(\lambda)| + \lambda |\beta_j|.$$

Pathwise coordinate optimization: Algorithm / 2

- This has an explicit solution with respect to β_j:

$$\tilde{\beta}_j(\lambda) = S\left(\sum_{i=1}^{N} x_{ij}(y_i - \sum_{k \neq j} x_{ik}\tilde{\beta}_k(\lambda)), \lambda\right),$$

  where

$$S(t, \lambda) = \text{sign}(t)(|t| - \lambda)_+$$

  is the soft-thresholding operator.
- Repeatedly iterating over the covariates results in the lasso estimate.

Degrees of freedom

$$\text{df}(\hat{\boldsymbol{y}}) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \text{Cov}(\hat{y}_i, y_i).$$

- For linear regression with k fixed predictors this gives: $\text{df}(\hat{\boldsymbol{y}}) = k$.
- For ridge regression: $\text{df}(\hat{\boldsymbol{y}}) = \text{tr}(\boldsymbol{H}_\lambda)$.
- For LAR after the kth step: $\text{df}(\hat{\boldsymbol{y}}) = k$.
- For lasso $\text{df}(\hat{\boldsymbol{y}})$ approximately equals the number of predictors in the model.
- For best subset selection if k variables are selected: $\text{df}(\hat{\boldsymbol{y}}) \geq k$.

Extensions

- Grouped Lasso:
  - For categorical variables the Lasso penalizes the individual dummy variables and selects them without taking into account that they belong to the same categorical variable.
  - Impose a penalty on the norm of the subvector of regression coefficients for the same categorical variable.
- Other penalties: e.g.,
  - SCAD (smoothly clipped absolute deviation) to achieve that larger coefficients are shrunken less.