# Penalized Regression

# Introduction

The results of regression methods are often not satisfactory due to lack of

- **Prediction accuracy:** Estimates of LMs and GLMs have low bias, but high variance. Prediction accuracy (which depends on the sum of variance and bias) can sometimes be improved by reducing the variance, while increasing the bias. A reduction in variance is achieved by shrinking estimates or setting them to 0, which increases the bias.
- **Interpretability:** With a large number of predictors, identifying a smaller subset exhibiting the strongest effects might be of interest.

# Best-subset or stepwise selection

- **Best-Subset:** Find for each $k \in \{0, 1, \ldots, p\}$ the subset of size $k$ that gives the best model fit (e.g. smallest RSS for a linear model).
  The leaps-and-bounds algorithm makes this feasible for $p$ as large as 30 or 40. See for example package **leaps**.
- **Stepwise**: A nested sequence of models is generated by adding or dropping terms depending on the search performed:
  - Forward
  - Backward
  - Hybrid

  Terms are added or removed as long as the performance criterion (e.g. AIC) is improved. See for example `step()` in the base distribution and `stepAIC()` in package **MASS**.

# OLS regression

The following optimization problem is solved in OLS regression

$$\begin{aligned}\hat{\beta}_{\text{OLS}} &= \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \right\} \\ &= \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\},\end{aligned}$$

i.e., the OLS estimator $\hat{\beta}_{\text{OLS}}$ minimizes the residual sum-of-squares (RSS).

The solution is given in closed form by

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

# Shrinkage

- 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:
  - Ridge
  - LASSO (least absolute shrinkage and selection operator)
  - Subset selection

# Penalized regression

In general the optimization problem in linear penalized regression is given by

$$\begin{aligned}\hat{\beta}_{\text{pen}} &= \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\phi(\beta) \right\} \\ &= \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\phi(\beta_1, \ldots, \beta_p) \right\},\end{aligned}$$

- The parameter $\lambda \geq 0$ is a complexity parameter, that controls the amount of shrinkage.
- The intercept $\beta_0$ is in general not shrunken.
- The function $\phi$ only depends on $\beta$, often on the "length" of $\beta$.
- This approach is sometimes also referred to as regularization.

# Ridge regression

The following optimization problem is solved in Ridge regression

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta_{[-0]}\|_2^2 \right\}$$
$$= \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

- The parameter $\lambda \geq 0$ is a complexity parameter, that controls the amount of shrinkage.
  - $\lambda = 0$: least squares fit.
  - $\lambda = \infty$: only a constant function $\beta_0$ fit.

# Ridge regression / 2

- An equivalent problem formulation is:

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}, \quad \text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t,$$

  with a one-to-one correspondence between $t$ and $\lambda$.
- The intercept $\beta_0$ is not shrunken.
- Ridge regression alleviates the problem of poorly identified coefficients in case of multicollinearity.
- The Ridge solutions are not equivariant under scaling of the regressors.
  $\Rightarrow$ One normally standardizes the regressors before analysis.

# Ridge regression / 3

- The Ridge criterion can also be written as

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^\top \beta,$$

with solution

$$\hat{\beta}_{\text{Ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y},$$

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

# Ridge: Shrinkage effects

- For orthonormal inputs:

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

- Otherwise use the singular value decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top,$$

with

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

- This gives

$$\mathbf{X}\hat{\beta}_{\text{ridge}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\top\mathbf{y}.$$

# Ridge: Shrinkage effects / 2

- Ridge regression
  - projects **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

$$\mathrm{df}(\lambda) = \mathrm{tr}[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top] = \mathrm{tr}(\mathbf{H}_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}.$$

# LASSO

The following optimization problem is solved in LASSO regression

$$\begin{aligned}\hat{\beta}_{\text{LASSO}} &= \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta_{[-0]}\|_1 \right\} \\ &= \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.\end{aligned}$$

This is equivalent to solving

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}, \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

# LASSO / 2

- LASSO is the abbreviation for "least absolute shrinkage and selection operator".
- The LASSO replaces the $L_2$ ridge penalty with the $L_1$ penalty.
- The parameter $\lambda \geq 0$ is a complexity parameter, that controls the amount of shrinkage.
- The intercept $\beta_0$ is not shrunken.
- 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 / 3

- The LASSO solutions are not equivariant under scaling of the regressors.
  $\Rightarrow$ One normally standardizes the regressors before analysis.
- The solution cannot be given in closed form, but efficient algorithms exist to compute the entire path of solutions as $\lambda$ is varied.

# 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:

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

  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:

$$\hat{\beta}_{\text{BSS}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}, \quad \text{subject to } \|\beta_{\setminus 0}\|_0 \leq t,$$

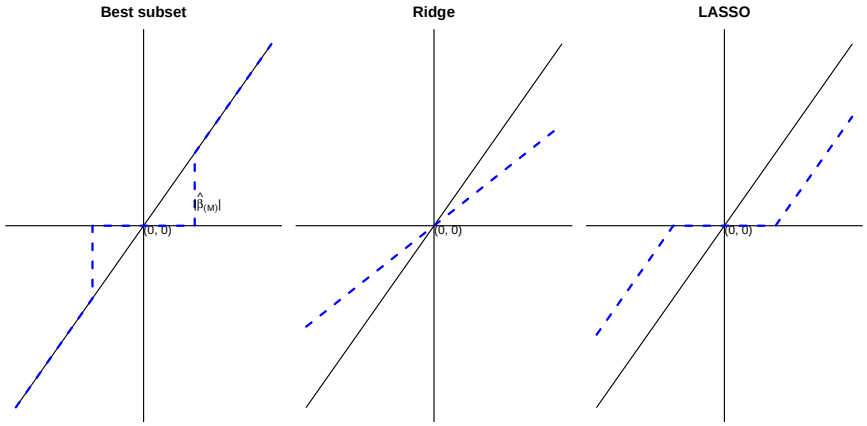
  with a one-to-one correspondence between $t$ and $\lambda$.

# Comparison

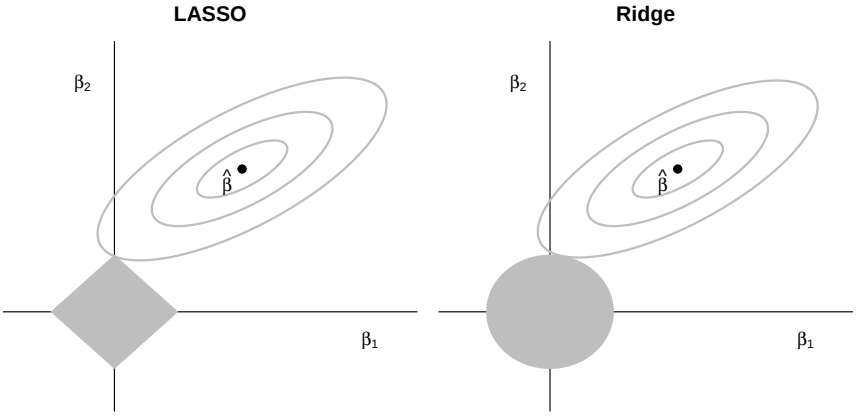
For orthonormal columns of the model matrix $\mathbf{X}$ the estimators are derived from the OLS estimates $\hat{\beta}_j$ by:

| Estimator | Formula |
|---|---|
| Best subset (size $M$) | $\hat{\beta}_j \cdot I(rank(\hat{\beta}_j) \leq M)$ |
| Ridge | $\hat{\beta}_j/(1+\lambda)$ |
| LASSO | $\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$ |

# Comparison / 2



# Comparison / 3



# Elastic net

The following optimization problem is solved in elastic net regression

$$\hat{\beta}_{\text{enet}} = \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda(\alpha\|\beta_{[-0]}\|_2^2 + (1-\alpha)\|\beta_{[-0]}\|_1) \right\}$$

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

- 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 $\beta_0$ is not shrunken.
- The elastic net solutions are not equivariant under scaling of the regressors.
  $\Rightarrow$ One normally standardizes the regressors before analysis.

# Shrinkage methods: Estimation

- Ridge:
  - Regression coefficient estimates are available in closed form for a given $\lambda$.
- LASSO:
  - If $\lambda$ 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 $\lambda$ values can be determined in a computationally efficient way using a similar algorithm as least angle regression (LAR).
  - Convex optimization problem (as ridge regression).

# Shrinkage methods: Estimation / 2

- 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 $A$.

# Degrees of freedom

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

- For linear regression with $p$ fixed predictors this gives: $\text{df}(\hat{\mathbf{y}}) = p$.
- For Ridge regression: $\text{df}(\hat{\mathbf{y}}) = \text{tr}(\mathbf{H}_\lambda)$.
- For LASSO $\text{df}(\hat{\mathbf{y}})$ approximately equals the number of predictors in the model.
- For best subset selection if $p$ variables are selected: $\text{df}(\hat{\mathbf{y}}) \geq p$.

# Selection of shrinkage parameters

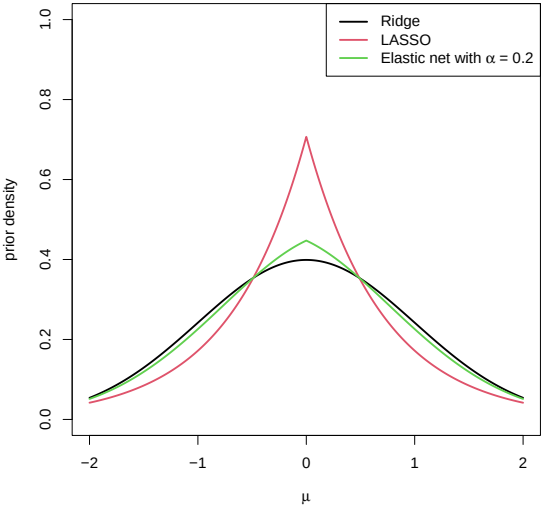
- **Testing / Training:**
  - Split data in testing and training data.
  - Select shrinkage parameters such that prediction performance is optimal in the training data.
- **$K$-fold cross-validation:**
  - Split data in $K$ subsets.
  - Fit model $K$ times where each time one subset is omitted.
  - Evaluate the performance of the $K$ models using the leave-out data.
  - Select the smallest model which does not perform significantly worse than the best performing model.

# 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 regularization approaches can be seen as determining the maximum a-posteriori estimates of the parameters using different prior distributions.

# Relation to Bayesian estimation / 2



# Software in R

Package **glmnet**:

- Fits a generalized linear model via penalized maximum likelihood.
- The regularization path is computed for the LASSO or elasticnet penalty at a grid of values for the regularization parameter $\lambda$.
- Implements $k$-fold cross-validation to select the hyperparameter $\lambda$.
- No formula interface is provided.
- By default an intercept is added and the covariates are standardized.

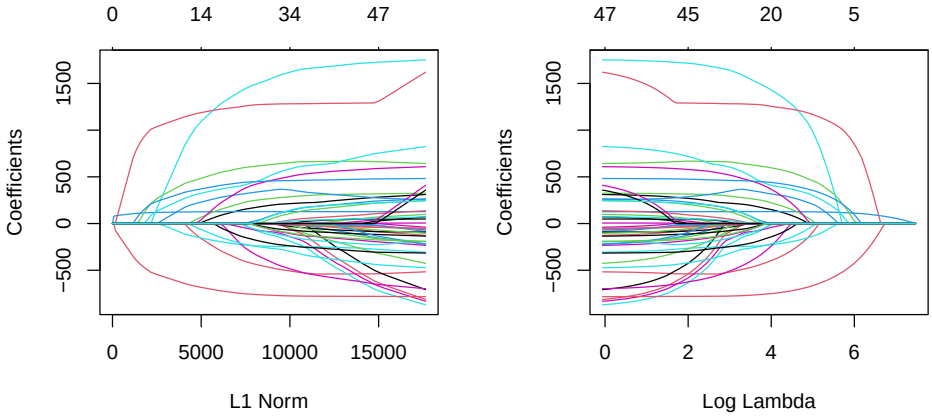
# Example: German Data Set

```
R> load("german.rda")
R> library("glmnet")
R> mf <- model.frame(Amount ~ . - Class, data = german)
R> X <- model.matrix(Amount ~ . - Class, data = mf)[, -1]
R> y <- model.response(mf)
```

LASSO:

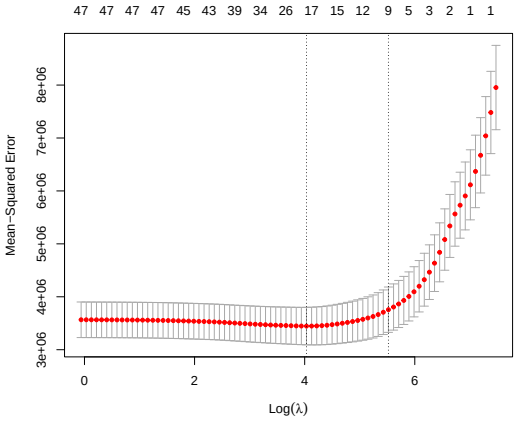
```
R> model <- glmnet(X, y)
```

# Example: German Data Set / 2




```
R> set.seed(1234)
R> cv.model <- cv.glmnet(X, y)
```

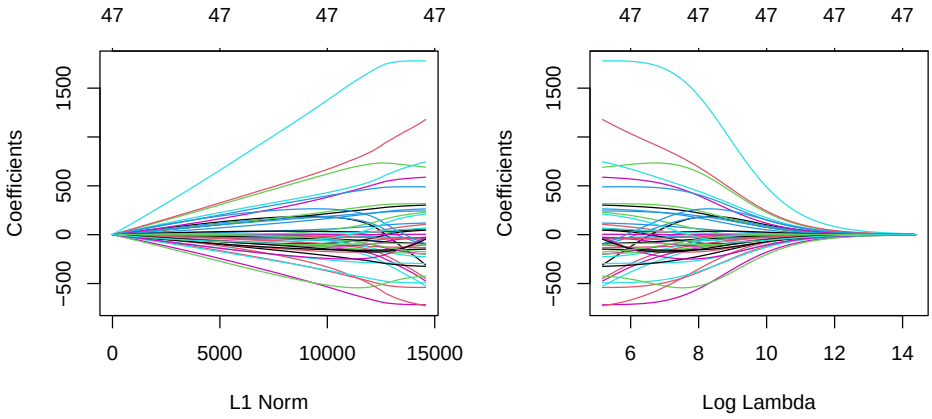
# Example: German Data Set / 3




Ridge:

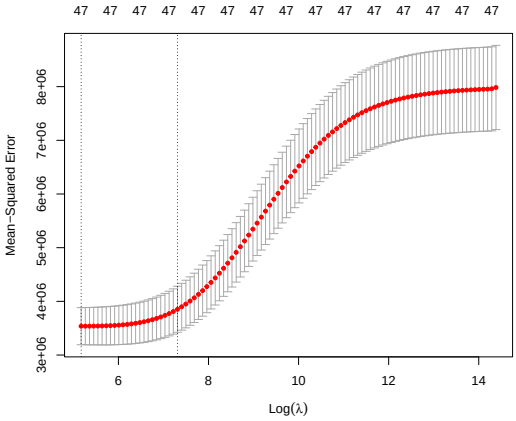
```
R> model <- glmnet(X, y, alpha = 0)
```

# Example: German Data Set / 4




```
R> set.seed(1234)
R> cv.model <- cv.glmnet(X, y, alpha = 0)
```

# Example: German Data Set / 5



# Penalized GLMs

In general the optimization problem in penalized GLMs is given by

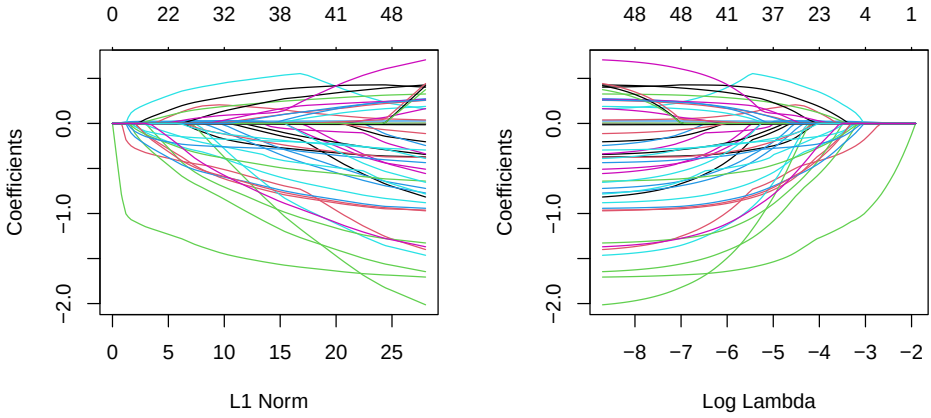
$$\hat{\beta}_{\text{pen}} = \arg\min_{\beta} \left\{ -2\ell(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) + \lambda\phi(\beta) \right\},$$

where twice the negative log-likelihood is penalized instead of the residual sum of squares.

# Example: German Data Set

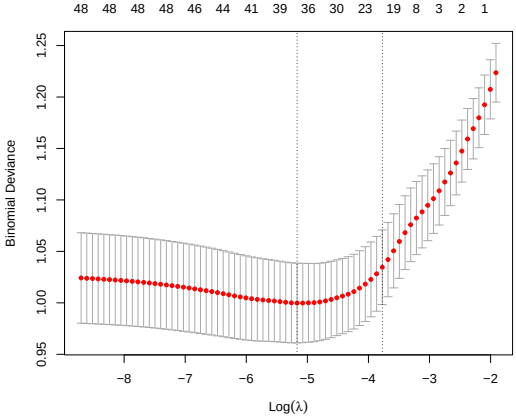
```
R> mf <- model.frame(Class ~ ., data = german)
R> X <- model.matrix(Class ~ ., data = mf)[, -1]
R> y <- model.response(mf)
R> model <- glmnet(X, y, family = "binomial")
```

# Example: German Data Set / 2




```
R> cv.model <- cv.glmnet(X, y, family = "binomial")
```

# Example: German Data Set / 3



# 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.