STATISTICAL LEARNING SYSTEMS

LECTURE 1: LINEAR REGRESSION MODELS REVISITED

Jacek Koronacki

Institute of Computer Science, Polish Academy of Sciences

Ph. D. Program 2013/2014

The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited

We are given n observations, each with p explanatory variables (predictors),

$$(x_{i1}, x_{i2}, \ldots, x_{ip}),$$

and one response variable, Y_i,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where ε_i are i.i.d. random errors with mean 0 and unknown variance σ^2, and $\beta_0, \ldots, \beta_p$ are unknown parameters.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited

In matrix notation,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_n)$,
$\boldsymbol{\varepsilon}' = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$,
$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$
and

$$\mathbf{X} = (x_{ij}) = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1,p} \\ 1 & x_{21} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{n,p} \end{pmatrix}.$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited

We estimate the unknown $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ by least squares (LS), i.e., by minimizing the residual sum of squares:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Upon differentiating w.r.t. $\boldsymbol{\beta}$ and equating the derivative to zero we get

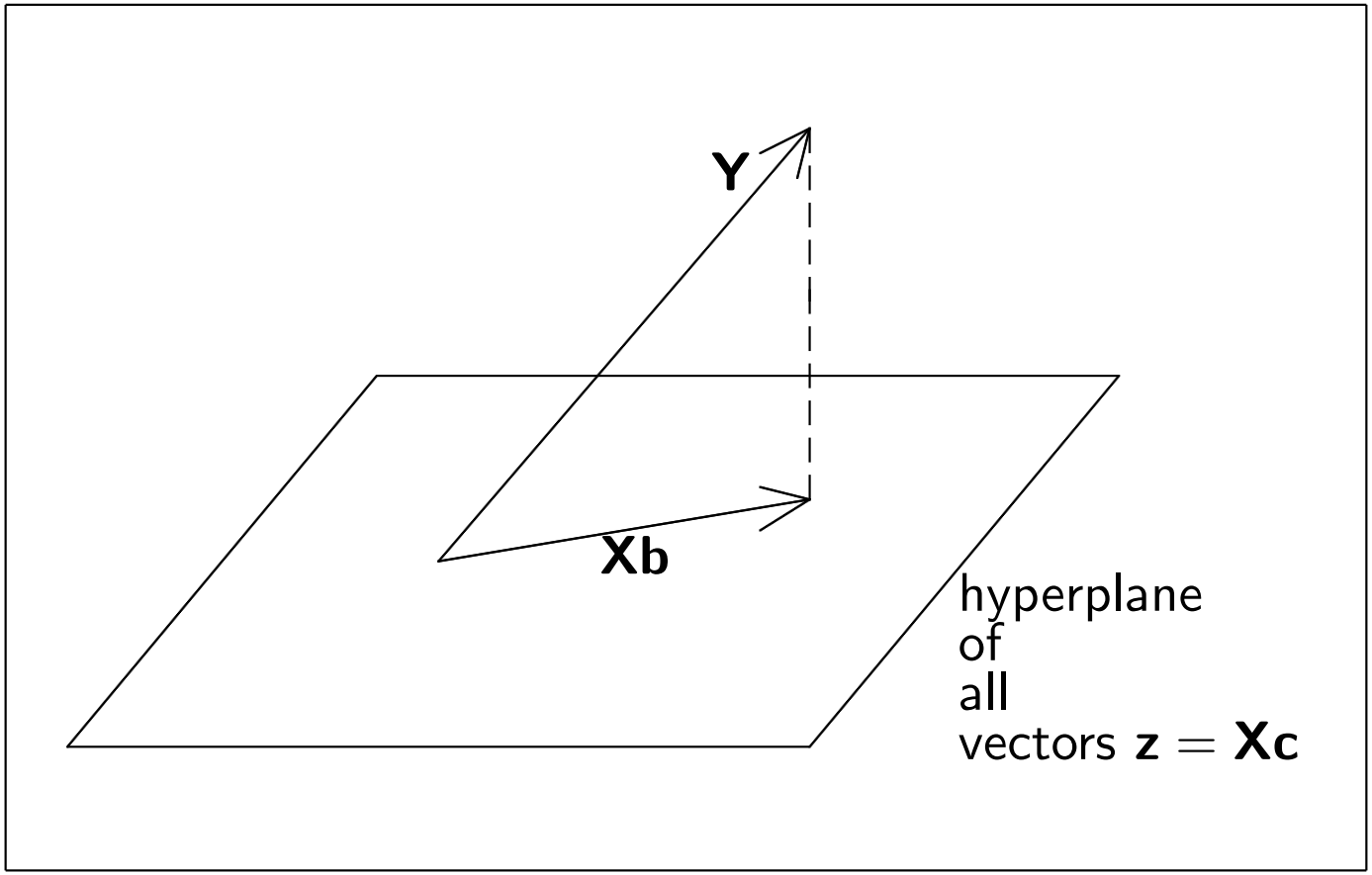
$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

and assuming that $\mathbf{X}$ has full rank, we obtain the unique solution

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited



The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - multiple regression from a univariate one

For a univariate model with no intercept, we have

$$b = \frac{< \mathbf{x}, \mathbf{y} >}{< \mathbf{x}, \mathbf{x} >},$$

$$\mathbf{r} = \mathbf{y} - \mathbf{x}b.$$

If the multiple inputs are orthogonal, then it is easy to see that

$$b_j = \frac{< \mathbf{x}_j, \mathbf{y} >}{< \mathbf{x}_j, \mathbf{x}_j >}$$

for each $j = 1, \ldots, p$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - multiple regression from a univariate one

Now, if we have a single input, but we have an intercept,

$$b_1 = \frac{< \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} >}{< \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} >}.$$

In turn, if there are multiple inputs which are not orthogonal, as is most often the case, we can orthogonalize them step by step:

- Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$.
- For $j = 1, 2, \ldots, p$
  - Regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} =< \mathbf{z}_\ell, \mathbf{x}_j > / < \mathbf{z}_\ell, \mathbf{z}_\ell >$, $\ell = 0, \ldots, j-1$ and residual $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$.
- Regress $\mathbf{y}$ on the residual $\mathbf{z}_p$ to give the estimate b_p.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - multiple regression from a univariate one

Let us restate the algorithm (known as the Gram-Schmidt procedure)

- Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$.
- For $j = 1, 2, \ldots, p$
  - Regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} =< \mathbf{z}_\ell, \mathbf{x}_j > / < \mathbf{z}_\ell, \mathbf{z}_\ell >$, $\ell = 0, \ldots, j-1$ and residual $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$.
- Regress $\mathbf{y}$ on the residual $\mathbf{z}_p$ to give the estimate b_p.

to note that each of the $\mathbf{x}_j$ is a linear combination of the $\mathbf{z}_k$, $k \leqslant j$, all the $\mathbf{z}_j$ are orthogonal (hence they form a basis for the column space of $\mathbf{X}$ and the LS projection onto this subspace is $\hat{\mathbf{y}}$), and $b_p =< \mathbf{z}_p, \mathbf{y} > / < \mathbf{z}_p, \mathbf{z}_p >$ is indeed the multiple regression coefficient of $\mathbf{y}$ on $\mathbf{x}_p$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - multiple regression from a univariate one

By rearranging the $\mathbf{x}_j$ we get that each jth multiple regression coefficient is the univariate regression coefficient of $\mathbf{y}$ on the residual after regressing $\mathbf{x}_j$ on $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p$.

We can write the 2nd step of the Gram-Schmidt procedure in matrix form:

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma},$$

where $\mathbf{Z}$ has as columns the $\mathbf{z}_j$ and $\mathbf{\Gamma}$ is the upper triangular matrix with entries $\hat{\gamma}_{kj}$. Introducing the diagonal matrix $\mathbf{D} = \mathrm{diag}\{||\mathbf{z}_0||, \ldots, ||\mathbf{z}_p||\}$, we get

$$\mathbf{X} = \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\mathbf{\Gamma} = \mathbf{Q}\mathbf{R},$$

the so-called *QR* decomposition of $\mathbf{X}$ with an $n \times (p+1)$ orthogonal $\mathbf{Q}$ and $(p+1) \times (p+1)$ upper triangular $\mathbf{R}$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - subset selection

Often needed to improve prediction accuracy (overfitting may occur, resulting in poor prediction accuracy by estimators which have low bias but large variance) and/or interpretation (given a large number of predictors we would like to determine a smaller subset which exhibits the strongest effect on the response).

- Best-subset selection (choose the smallest model that minimizes an estimate of the expected prediction error)
- Forward-stepwise selection (start with the intercept, then add the predictor that most improves the fit; $n > p$)
- Backward-stepwise selection (delete the predictor that has the least impact on the fit)
- Forwad-stagewise regression (center predictors and start with the intercept equal to $\bar{y}$; then identify the predictor most correlated with the current residual; $n > p$).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - shrinkage methods

Another way out of trouble is via regularization.

Ridge regression (henceforth we assume that the inputs are centered and $\hat{\beta}_0$ is estimated by $\bar{y}$):

$$\hat{\beta}^{\mathrm{RIDGE}} = \mathrm{argmin}_{\beta} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda > 0$ is a parameter, and $\lambda \sum_{j=1}^{p} \beta_j^2$ is the penalty for large magnitude of coefficients. The above is equivalent to (and is the Lagrangian for) the minimization of

$$\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 \leqslant s \quad \text{for some } s = s(\lambda).$$

Linear regression models revisited - shrinkage methods

Writing the penalized residual sum of squares in matrix form,

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta'\beta$$

one easily obtains

$$\hat{\beta}^{\mathrm{RIDGE}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}.$$

- That is, we add λ to the diagonal of $\mathbf{X}'\mathbf{X}$ before inverting it (what helps when $\mathbf{X}'\mathbf{X}$ is almost singular i.e. explanatory variables are approx. linearly dependent (collinear).
- One needs to standardize explanatory variables before applying ridge regression.
- Parameter λ is usually chosen by cross-validation.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - shrinkage methods

The ridge approach makes $\hat{y} = (\hat{\beta}^{\mathrm{RIDGE}})'\mathbf{x}$ more sensitive to directions in which the $\mathbf{x}_i$ are varying most.

To see this recall that by the singular value decomposition (SVD) any $n \times p$ matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

where $\mathbf{U}$ is an $n \times p$ column-orthogonal matrix and $\mathbf{V}$ is a $p \times p$ orthogonal matrix, with the columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$, and the columns of $\mathbf{V}$ spanning the row space; $\mathbf{D} = \mathrm{diag}\{\mathrm{d}_1, \ldots, \mathrm{d}_\mathrm{p}\}$, $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited

Now, we can write the LS fitted vector as

$$\mathbf{Xb} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{U}\mathbf{U}'\mathbf{y},$$

and the ridge solutions as

$$\mathbf{X}\hat{\beta}^{\mathrm{RIDGE}} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}'\mathbf{y}$$

and, finally,

$$\mathbf{X}\hat{\beta}^{\mathrm{RIDGE}} = \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j'\mathbf{y}.$$

Regarding shrinking, the question remains what it means that a d_j^2 is small.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

For the centered $\mathbf{X}$, the sample covariance matrix is $\mathbf{S} = \mathbf{X}'\mathbf{X}/n$.
We have

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}'$$

which is the eigendecomposition of $\mathbf{X}'\mathbf{X}$.

It is well known (we shall come back to it in detail later) that the columns of $\mathbf{V}$, $\mathbf{v}_j$ are the so-called principal components of $\mathbf{X}$ and, more importantly, that the 1st principal component direction $\mathbf{v}_1$ has the property that $\mathbf{z}_1 = \mathbf{X}\mathbf{v}_1$ has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$, the 2nd principal component direction $\mathbf{v}_2$ has the property that $\mathbf{z}_2 = \mathbf{X}\mathbf{v}_2$ has largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$ which are orthogonal to the 1st principal component direction, etc.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - shrinkage methods

It is easily seen that $\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = \mathbf{u}_j d_j$ and that the sample variance of $\mathbf{z}_j$ is equal to d_j^2/n.

Hence the small singular values d_j correspond to directions in the column space of $\mathbf{X}$ which have small variance, and ridge regression shrinks these directions the most. (Recall that

$$\mathbf{X}\hat{\beta}^{\mathrm{RIDGE}} = \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j' \mathbf{y}.)$$

We call

$$\mathrm{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}'] = \sum_{\mathrm{j}=1}^{\mathrm{p}} \frac{\mathrm{d}_\mathrm{j}^2}{\mathrm{d}_\mathrm{j}^2 + \lambda}$$

the effective degrees of freedom of the ridge regression fit.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

The Lasso:

$$\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leqslant t.$$

The Lasso, in contrast to ridge regression, eliminates for small t some variables from the model. It can thus be used as a feature selection method.

It happens that the least angle regression (LAR), to be discussed in a moment, is not only intimately connected with the lasso, but also provides an efficient algorithm for computing the entire lasso path (for all values of t).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - shrinkage methods

However, before introducing LAR, let us present a simple comparison of the following approaches to restricting the linear regression model: subset selection, ridge regression and the lasso. In the discussion below we confine ourselves to the case of an orthonormal columns of $\mathbf{X}$:

- For the best subset selection (of size M), the LS $\hat{\beta}_j$ are transformed to $\hat{\beta}_j \cdot I(|\hat{\beta}_j| \geqslant |\hat{\beta}_{(M)}|)$
- For ridge regression, they are transformed to $\hat{\beta}_j/(1+\lambda)$
- And for the lasso, to $\mathrm{sign}(\hat{\beta}_\mathrm{j})(|\hat{\beta}_\mathrm{j}| - \lambda)_+$.

Introducing the elastic-net penalty

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

with, say, $\alpha = 0.2$, leads to some compromise between ridge and lasso.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Least Angle Regression (LAR):

- Standardize the predictors. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.
- Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
- Move β_j from 0 towards its LS coefficient $< \mathbf{x}_j, \mathbf{r} >$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
- Move β_j and β_k in the direction defined by their joint LS coefficient of the current residual on $\mathbf{x}_j, \mathbf{x}_k$, until some other $\mathbf{x}_\ell$ has as much correlation with the current residual.
- Continue until all p predictors have been entered. After $\min(n-1, p)$ steps, we arrive at the full LS solution.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Suppose $\mathcal{A}_k$ is the active set of predictors at the beginning of the kth step, and let $\boldsymbol{\beta}_{\mathcal{A}_k}$ be the coefficient vector for these predictors. If $\mathbf{r}_k = \mathbf{y} - \mathbf{X}_{\mathcal{A}_k}\boldsymbol{\beta}_{\mathcal{A}_k}$ is the current residual, then the direction for this step is

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

The coefficient profile then evolves as $\boldsymbol{\beta}_{\mathcal{A}_k}(\alpha) = \boldsymbol{\beta}_{\mathcal{A}_k} + \alpha\boldsymbol{\delta}_k$. It can be shown that this keeps the correlations tied and decreasing.

If the fit vector at the beginning of this step is $\hat{\mathbf{f}}_k$, then it evolves as

$$\hat{\mathbf{f}}_k(\alpha) = \hat{\mathbf{f}}_k + \alpha\mathbf{u}_k,$$

where $\mathbf{u}_k = \mathbf{X}_{\mathcal{A}_k}\boldsymbol{\delta}_k$ is the new fit direction (hence the name of the method).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - shrinkage methods

It has been observed that the lasso and LAR coefficient profiles are almost identical until for the first time one coefficient passes back through zero.

Thus, the following modification (the so-called lasso modification) has been proposed to the LAR algorithm, which gives the entire lasso path (both the lasso and LAR coefficient profiles are piecewise linear and for the latter the exact step length can be calculated at the beginning of each step):

LAR(lasso) modification:
If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint LS direction.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - shrinkage methods

Heuristics behind the similarity between the lasso and the LAR:

Suppose $\mathcal{A}$ is the active set of variables for the LAR, tied in their absolute inner-product with the current residuals $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$:

$$\mathbf{x}_j'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \gamma \cdot s_j$$

for each $j \in \mathcal{A}$, where $s_j \in \{-1, 1\}$. Note that $|\mathbf{x}_k'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})| \leqslant \gamma$ for each $k \notin \mathcal{A}$. Now consider the lasso criterion,

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

Let $\mathcal{B}$ be the active set of variables in the solution for a given λ. The stationarity conditions for $R(\boldsymbol{\beta})$ give

$$\mathbf{x}_j'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \lambda \cdot \mathrm{sign}(\boldsymbol{\beta}_\mathrm{j})$$

for each $j \in \mathcal{B}$, in full analogy to the LAR solution as long as the sings match.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - shrinkage methods

Before we present another LAR-like algorithm, let us discuss the Incremental Forward Stagewise Regression algorithm (FS_ε):

- Standardize all the predictors and start with residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$ and $\beta_1 = \cdots = \beta_p = 0$.
- Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
- Update $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j = \varepsilon \cdot \mathrm{sign}[< \mathbf{x}_\mathrm{j}, \mathbf{r} >] and \varepsilon$ is a small step size, and set $\mathbf{r} \leftarrow \mathbf{r} - \delta_j \mathbf{x}_j$.
- Repeat steps 2 and 3, until the residuals are uncorrelated with all the predictors.

Let $\varepsilon \rightarrow 0$ to obtain the Infinitesimal Forward Stagewise Regression (FS_0).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Linear regression models revisited - shrinkage methods

The following modification of the LAR algorithm implements FS_0:

- Find the new direction by solving the constrained LS problem

$$\min_{\mathbf{b}} ||\mathbf{r} - \mathbf{X}_{\mathcal{A}}\mathbf{b}||_2^2$$

subject to $b_j s_j \geqslant 0$ for all $j \in \mathcal{A}$, where s_j is the sign of $< \mathbf{x}_j, \mathbf{r} >$.

That is, the modification consists in replacing the LS fit by a non-negative LS fit. (Recall that steps 4 and 5 of the original LAR read: 4. move β_j and β_k in the direction defined by their joint LS coefficient of the current residual on $\mathbf{x}_j, \mathbf{x}_k$, until some other $\mathbf{x}_\ell$ has as much correlation with the current residual. 5. Continue until all p predictors have been entered.)

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

It can be shown that if the LAR profiles are monotone, then all three methods - LAR, lasso, and FS_0 - give identical profiles. If the profiles are not monotone but do not cross the zero axis, then LAR and lasso are identical.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund