# Statistical learning for biological data

David Causeur
*Department of Statistics and Computer Science*
*Agrocampus Ouest*
*IRMAR CNRS UMR 6625*
*http://www.agrocampus-ouest.fr/math/causeur/*

# Outline

1 Regression modeling
   Why ’regression’?
   Fitting linear regression models
   Feature selection for prediction

2 Penalized regression
   Sparse regression modeling
   Penalized estimation of classification models

# Outline

1. Regression modeling
   - Why 'regression'?
   - Fitting linear regression models
   - Feature selection for prediction
2. Penalized regression
   - Sparse regression modeling
   - Penalized estimation of classification models

# Impact of correlation on the least-squares fit

Case $p = 1$: $\mathbb{E}_x(Y) = \beta_0 + \beta x$

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{n}\frac{1}{S_x^2}, \text{ where } S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Estimation accuracy depends on the spread of the $x-$profile.

# Impact of correlation on the least-squares fit

Case $p = 2$: $\mathbb{E}_x(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{n}\begin{bmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{bmatrix}^{-1} = \begin{bmatrix} \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{Var}(\hat{\beta}_2) \end{bmatrix}$$

Hence,

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n}\frac{1}{S_1^2}\frac{1}{1 - r_{12}^2}, \quad \text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{n}\frac{1}{S_2^2}\frac{1}{1 - r_{12}^2}$$

Estimation accuracy also depends on the correlation (here $r_{12}$) within the $x-$profile.

▶ Illustration using high-dimensional data in R

# Impact of correlation on the least-squares fit

## Ordinary Least Squares fitting

$$\hat{\beta} \text{ minimizes} \sum_{i=1}^{n} \Big( Y_i - \bar{Y} - \left[ \beta_1 (x_{i1} - \bar{x}_1) + \ldots + \beta_p (x_{ip} - \bar{x}_p) \right] \Big)^2$$

## A closed-form solution ... provided $S_x^{-1}$ exists

$$\hat{\beta} = S_x^{-1} s_{xy}.$$

$\hat{\beta}$ is unbiased with variance

$$\mathrm{Var}_x(\hat{\beta}) = \frac{\sigma^2}{n} S_x^{-1}.$$

In high-dimension, $\mathrm{Var}_x(\hat{\beta}_j)$ can be very large ...

# A biased alternative: penalized regression

## Least-squares optimization under constraint

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n}\Big(Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \ldots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}]\Big)^2,$$

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

# A biased alternative: penalized regression

## Least-squares optimization under constraint (equivalent form)

$$\text{RSS}(\beta) = \sum_{i=1}^{n}\big(Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \ldots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}]\big)^2,$$

$$\text{subject to } \sum_{j=1}^{p}\beta_j^2 = \kappa$$

# A biased alternative: penalized regression

## Least-squares optimization under constraint (equivalent form)

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n}\left(Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \ldots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}]\right)^2,$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = \kappa$$

## The *Lagrange multiplier* trick:

$$\mathrm{RSS}(\beta;\lambda) = \sum_{i=1}^{n}\left(Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \ldots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}]\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

# A biased alternative: penalized regression

Let's do it when $p = 1$:

$$\begin{aligned}
\mathrm{RSS}(\beta; \lambda) &= \sum_{i=1}^{n} \left( Y_i - \bar{Y} - \beta[x_i - \bar{x}] \right)^2 + \lambda\beta^2, \\
&= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - 2\beta \sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x}) \\
&\quad + \beta^2 \left[ \lambda + \sum_{i=1}^{n} (x_i - \bar{x})^2 \right], \\
&= n\left( s_Y^2 - 2\beta s_{xY} + \beta^2 \left[ s_x^2 + \frac{\lambda}{n} \right] \right).
\end{aligned}$$

# A biased alternative: penalized regression

Let's do it when $p = 1$:

$$\mathrm{RSS}(\beta; \lambda) \quad = \quad n\big(s_Y^2 - 2\beta s_{xY} + \beta^2 \big[s_x^2 + \frac{\lambda}{n}\big]\big)$$

Let's differentiate w.r.t $\beta$:

$$\frac{\partial \mathrm{RSS}}{\partial \beta}(\beta; \lambda) \quad = \quad -2n\big(s_{xY} - \beta\big[s_x^2 + \frac{\lambda}{n}\big]\big).$$

By equating to zero:

$$\hat{\beta}_\lambda = \frac{s_{xY}}{s_x^2 + \frac{\lambda}{n}}$$

# A biased alternative: penalized regression

Ridge estimation of $\beta$:

$$\hat{\beta}_\lambda = \left[ S_{xx} + \frac{\lambda}{n} I_p \right]^{-1} s_{xy}$$

The ridge estimator

- is biased: $b_\lambda = \mathbb{E}_x \left[ \hat{\beta}_\lambda - \beta \right] \neq 0$ [shrinkage estimation]
- but has a smaller variance than $\hat{\beta}$:
  - $\lambda \approx 0$: small bias, large variance
  - $\lambda \gg 0$: large bias, small variance

$\lambda$ is chosen so that MSEP($\lambda$) is as small as possible

▶ Ridge regression using R

# Shrinkage and selection: *LASSO* regression

Least-squares optimization under constraint

$$
\begin{aligned}
\mathrm{RSS}(\beta;\lambda) &= \sum_{i=1}^{n}\left(Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \ldots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}]\right)^2 \\
&+ \lambda \sum_{j=1}^{p} |\beta_j|
\end{aligned}
$$

# Shrinkage and selection: *LASSO* regression

Let's try to do it when $p = 1$:

$$
\begin{aligned}
\mathrm{RSS}(\beta;\lambda) &= \sum_{i=1}^{n}\big(Y_i - \bar{Y} - \beta[x_i - \bar{x}]\big)^2 + \lambda|\beta|, \\
&= n\Big(s_Y^2 - 2\beta s_{xY} + \beta^2 s_x^2 + \frac{\lambda}{n}|\beta|\Big), \\
&= n\begin{cases} s_Y^2 - 2\beta[s_{xY} + \frac{\lambda}{2n}] + \beta^2 s_x^2 & \text{for} \quad \beta < 0, \\ s_Y^2 - 2\beta[s_{xY} - \frac{\lambda}{2n}] + \beta^2 s_x^2 & \text{for} \quad \beta \geq 0 \end{cases}
\end{aligned}
$$

# Shrinkage and selection: *LASSO* regression

Let's try to do it when $p = 1$:

$$\text{RSS}(\beta;\lambda) = n \begin{cases} s_Y^2 - 2\beta[s_{xY} + \frac{\lambda}{2n}] + \beta^2 s_x^2 & \text{for} \quad \beta < 0, \\ s_Y^2 - 2\beta[s_{xY} - \frac{\lambda}{2n}] + \beta^2 s_x^2 & \text{for} \quad \beta \geq 0 \end{cases}$$

Not differentiable w.r.t $\beta$ for $\beta = 0$...

$$\hat{\beta}_\lambda = \begin{cases} \frac{s_{xY} + \frac{\lambda}{2n}}{s_x^2} & \text{if} \quad s_{xY} \leq -\frac{\lambda}{2n}, \\ 0 & \text{if} \quad -\frac{\lambda}{2n} \leq s_{xY} \leq \frac{\lambda}{2n}, \\ \frac{s_{xY} - \frac{\lambda}{2n}}{s_x^2} & \text{if} \quad s_{xY} \geq \frac{\lambda}{2n} \end{cases}$$

- Implementation using R

# Shrinkage and selection: *LASSO* regression

Extension to a multivariate profile of *xs*: no closed-form expression for $\hat{\beta}_\lambda$

... estimation achieved using an iterative algorithm (e.g. cyclic coordinate descent)

The LASSO estimator is also biased

- $b_\lambda = \mathbb{E}_x \left[ \hat{\beta}_\lambda - \beta \right] \neq 0$ [shrinkage estimation]
- ... but $\hat{\beta}_\lambda$ has a smaller variance than $\hat{\beta}$
- ... and increasing $\lambda$ kills the $\beta$s [selection]

▶ LASSO regression using R

# LASSO or Ridge?

Both can be very performant

LASSO regression models are sparse

Warning!! When the *x*s are highly correlated, the selection is unstable ...

Mixing ridge and lasso penalties (elastic net) may bring stability:

$$\hat{\beta}(\lambda, \alpha) \text{ minimizes } \mathrm{RSS}(\beta) + \lambda \left[ \alpha \sum_{j=1}^{p} |\beta_j| + (1 - \alpha) \sum_{j=1}^{p} \beta_j^2 \right]$$

Warning!! Two hyper-parameters to be tuned ...

# Ridge/LASSO logistic regression modeling

Aim: guessing production site of coffees using NIRS

- $Y$: production site named $y_1, \dots, y_6$
- $x$ is a NIRS:

$$x = \big(x(\omega_1), \dots, x(\omega_p)\big)', \text{ where } \omega_i \text{ is the } i\text{th wave number.}$$

▶ Import data in R session

# Ridge/LASSO logistic regression modeling

## Minimization of the penalized deviance

$$\begin{aligned}\mathcal{D}(\beta;\lambda) &= \mathcal{D}(\beta)+\lambda||\beta||_2^2, \quad \text{[Ridge]}\\ \mathcal{D}(\beta;\lambda) &= \mathcal{D}(\beta)+\lambda||\beta||_1, \quad \text{[Lasso]}\end{aligned}$$

where $||\beta||_2^2 = \sum_{k=2}^{6}\sum_{j=1}^{p}[\beta_j^{(k)}]^2$ and $||\beta||_1 = \sum_{k=2}^{6}\sum_{j=1}^{p}|\beta_j^{(k)}|$.

► Lasso estimation of a multinomial logistic regression model using `R`