Introduction to Model Selection and Regularization

Ghislain Durif, Laurent Modolo, Franck Picard

Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon

franck.picard@ens-lyon.fr

Outline

1. **The multi-loci model, algebra of linear models**
2. Increasing the dimension of linear models
3. The bias-variance trade-off
4. Generalization error and cross validation
5. Regularization and Penalization
6. Feature selection

General Model with many loci

- GWAS provide millions SNPs: model the impact of all SNPs on a phenotype.
- The one-locus model :

$$\mathbb{E}(Y_{ij} \mid x_{ij}) = \mu_i x_{ij}$$

- The multilocus model without interaction/epistasis

$$\mathbb{E}(Y_{ij} \mid x_{ij}) = \sum_{\ell=1}^{L} \mu_i^\ell x_{ij}^\ell$$

- The multilocus model with interaction/epistasis of order 1

$$\mathbb{E}(Y_{ij} \mid x_{ij}) = \sum_{ij\ell} \mu_i^\ell x_{ij}^\ell + \sum_{ij\ell} \sum_{i'j\ell'} \gamma_{i,i'}^{\ell\ell'} x_{ij}^\ell \times x_{i'j}^{\ell'}$$

- etc ...
- The model could be complexified, at what cost ?

General presentation of linear models

- Let us consider $\mathbf{Y}$ a response vector of size n, and $\mathbf{X}$ a matrix of regressors, of size $n \times p$
- The entries of $\mathbf{X}$ can be binary (ANOVA) or continuous (regression) or both
- p is the total number of regressors (features)
- We consider the linear model :

$$\mathbb{E}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

- The relationship between the response and regressors is a linear combination of parameters $\boldsymbol{\beta} \in \mathbb{R}^p$.

The least-square method (LS)

- The estimation of $\boldsymbol{\beta}$ consists in finding an estimator $\widehat{\boldsymbol{\beta}}$ so that the model $\mathbb{E}(\mathbf{Y} \mid \mathbf{X})$ is not too far from the observations

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \underset{\beta \in \mathbb{R}^p}{\arg\min}\, d^2\big(\mathbf{Y}, \mathbb{E}(\mathbf{Y} \mid \mathbf{X})\big) \\
&= \underset{\beta \in \mathbb{R}^p}{\arg\min}\, d^2\big(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}\big) \\
\widehat{\boldsymbol{\beta}} &= \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2
\end{aligned}$$

- The LS estimator of $\boldsymbol{\beta}$ minimizes the euclidean distance between the observations and the model.
- The best solution is $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, the best linear predictor of $\mathbf{Y}$ based on $\mathbf{X}$.

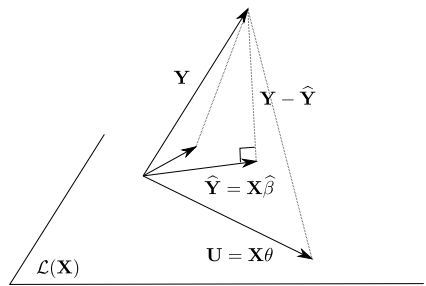
Model Space

- Linear model: $\mathbb{E}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$
- Solution are a linear combination of features

$$\mathbf{U} = \mathbf{X}\boldsymbol{\theta}$$

- $\mathcal{L}(\mathbf{X}) = \text{Span}(\mathbf{x}^1, \ldots, \mathbf{x}^p)$, the vector space spanned by the columns of $\mathbf{X}$

$$\forall \mathbf{U} \in \mathcal{L}(\mathbf{X}), \quad \mathbf{U} = \mathbf{X}\boldsymbol{\theta}$$



Orthogonal projection (1)

- Consider $\mathbf{Y} - \widehat{\mathbf{Y}}$, the residuals (error term) of the model
- The minimizer $\widehat{\mathbf{Y}}$ is of the form:

$$\|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2 \quad = \quad \min_{\mathbf{U} \in \mathcal{L}(\mathbf{X})} \|\mathbf{Y} - \mathbf{U}\|_2^2$$

- This minimization has an explicit solution: the orthogonal projection of $\mathbf{Y}$ onto $\mathcal{L}(\mathbf{X})$
- If we consider $\mathbf{P}_{n \times n}$ the orthogonal projector onto $\mathcal{L}(\mathbf{X})$, such that

$$\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}, \quad \mathbf{P}' = \mathbf{P}, \quad \mathbf{P}^2 = \mathbf{P}$$

- How can we derive the orthogonal projector in terms of $\mathbf{X}$?

Orthogonal projection (2)

- The idea is that the residuals $\mathbf{Y} - \widehat{\mathbf{Y}}$ should be orthogonal to $\mathcal{L}(\mathbf{X})$

$$\forall j \in \{1, \ldots, p\}, \quad \langle \mathbf{x}^j, \mathbf{Y} - \widehat{\mathbf{Y}} \rangle = 0$$

- Which provides the normal equations

$$\mathbf{X}' \left(\mathbf{Y} - \widehat{\mathbf{Y}} \right) = \mathbf{X}' \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right) = 0$$

- Suppose that $n \gg p$, and $\text{rank}(\mathbf{X}) = p$, we have

$$\mathbf{P} = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Y}$$

Statistical interpretation of the LS-estimator

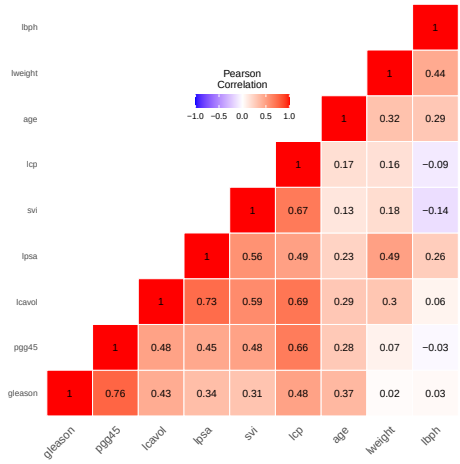
- The least-square estimator is a projection estimator
- $\mathbf{S} = \mathbf{X}'\mathbf{X}$ is the empirical covariance of regressors (including their dependencies)

$$\widehat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{X}'\mathbf{Y}$$

- $\widehat{\boldsymbol{\beta}}$: generalized correlation coefficient normalized by the dependencies between regressors
- If features are highly correlated (redundancy), $\mathbf{S}^{-1}$ is degenerate
- The feature-based (marginal) analysis vs. global analysis that accounts for dependencies between regressors.

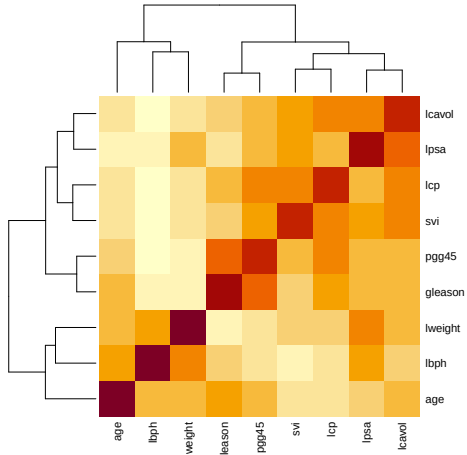
Example of correlation between features

- Prostate Cancer data [**?**], $n = 97$, $p = 10$
- Predict PSA levels using clinical covariates
- High correlation between features
- Redundancy between features



Example of correlation between features

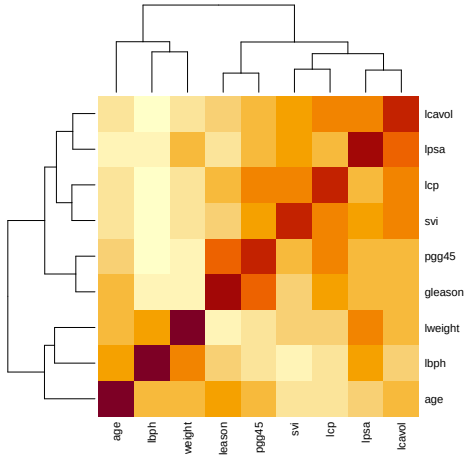
- Prostate Cancer data [**?**], $n = 97$, $p = 10$
- Predict PSA levels using clinical covariates
- High correlation between features
- Redundancy between features



Example of correlation between features

```
summary(model.full)
[1] 0.5221043
Call:
lm(formula = lpsa ~ ., data = prostate.train)
Residuals:
     Min       1Q   Median       3Q      Max
-1.64870 -0.34147 -0.05424  0.44941  1.48675
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.4292     1.5536   0.276  0.78334
lcavol         1.0466     0.1950   5.366 1.47e-06 ***
lweight        2.2623     0.8224   2.751  0.00792 **
age           -1.2477     0.8938  -1.396  0.16806
lbph           0.2123     0.1032   2.056  0.04431 *
svi            0.3515     0.1423   2.469  0.01651 *
lcp           -0.2924     0.1566  -1.867  0.06697 .
gleason       -0.2012     1.3716  -0.147  0.88389
pgg45          0.3737     0.2151   1.738  0.08755 .
```

Some coefficients have show variance



Outline

1. The multi-loci model, algebra of linear models
2. **Increasing the dimension of linear models**
3. The bias-variance trade-off
4. Generalization error and cross validation
5. Regularization and Penalization
6. Feature selection

Prediction errors when p increases

- What is the impact of dimensions on the prediction errors

$$\frac{1}{n}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2 = \mathcal{O}\left(\frac{p}{n}\right)$$

- If we add new individuals : $n \to \infty$ and $p \ll n$

$$\frac{1}{n}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2 \to 0$$

- If we add new features : $p \to \infty$ and $p \gg n$

$$\frac{1}{n}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2 = \mathcal{O}(1)$$

Interpretation of the features contribution

- Can we determine the set of contributing features ?

$$\forall j \in \{1, \dots, p\}, \quad \widehat{\beta}_j \neq 0$$

- We can perform a test followed by multiple testing

$$\mathcal{H}_0^j : \quad \{\beta_j = 0\}$$

- Use the LS-estimator, note the marginal strategy
- We can suppose that among p features, only s^* are non null

$$\{1, \dots, p\} = \mathcal{S}_0 \cup \mathcal{S}_1$$

$$\mathcal{S}_0 : \{j,\, \beta_j^* = 0\}, \quad \mathcal{S}_1 : \{j,\, \beta_j^* \neq 0\}$$

Sparsity assumption

- This hypothesis can be restated with the norm of the parameter

$$\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} \mathbf{1}_{\{\beta_j \neq 0\}} = s^*$$

- Then if we suppose that the number of non-null parameters $s^* \ll n$

$$\frac{1}{n}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2 = \mathcal{O}\left(\frac{s^*}{n}\right) \longrightarrow 0$$

- Unfortunately, $\mathcal{S}_1$ is unknown, with complexity $\mathcal{C}_n^p$
- The sparsity assumption has consequences on the optimization algorithm to find $\widehat{\boldsymbol{\beta}}$

Constrained optimization

- The LS-estimator is the solution of:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

- The constrained estimator:

$$\widehat{\boldsymbol{\beta}}_s = \underset{\|\boldsymbol{\beta}\|_0 \leq s}{\arg\min} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

- This can be reformulated in

$$\widehat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0$$

- λ is a hyper parameter to balance the intensity of the constraint, to be tuned

Penalized estimators

- The approach can be generalized:

$$\widehat{\boldsymbol{\beta}}_\lambda \quad = \quad \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \,\text{pen}(\boldsymbol{\beta})$$

- $\text{pen}(\boldsymbol{\beta})$ is a penalty function that scores the complexity of the model
- The LASSO estimator penalizes by the intensity of the coefficients

$$\text{pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$$

- The RIDGE estimator penalizes by the square of the coefficients

$$\text{pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$$

Outline

1. The multi-loci model, algebra of linear models
2. Increasing the dimension of linear models
3. **The bias-variance trade-off**
4. Generalization error and cross validation
5. Regularization and Penalization
6. Feature selection

General presentation

- We have some data recording on the same individuals $(Y_i, X_i)_{i=1,n}$
- These data are characterized by a joint distribution $\mathbb{P}(\mathbf{Y}, \mathbf{X})$
- We suppose that $\mathbf{Y}$ is the response variable, $\mathbf{X}$ are the features, supposed fixed

$$\mathbb{P}(\mathbf{Y}, \mathbf{X}) = \mathbb{P}(\mathbf{Y} \mid \mathbf{X})\mathbb{P}(\mathbf{X})$$

- To learn the relationship between $\mathbf{Y}$ and $\mathbf{X}$ we suppose that:

$$Y_i = f(X_i) + \varepsilon_i$$

- f: a model to links the features to the response (unknown)
- ε_i: random prediction errors, centered $\mathbb{E}(\varepsilon_i) = 0$, often supposed $i.i.d.$

What is f ?

- We want to minimize the prediction error by finding f such that

$$f = \arg\min_{\varphi} \mathbb{E}\Big[\Big(\mathbf{Y} - \varphi(\mathbf{X})\Big)^2\Big]$$

- The best prediction of $\mathbf{Y}$ at every point $\mathbf{X}$ in terms of quadratic risk is the conditional expectation:

$$f(\mathbf{X}) = \mathbb{E}\big(\mathbf{Y} \mid \mathbf{X}\big)$$

- In linear regression we suppose that the conditional expectation is linear wrt $\mathbf{X}$

$$f(\mathbf{X}; \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$$

- Then the determination of the model consists in estimating $\boldsymbol{\beta}$ (parametric model)

What is a model collection ?

- We define a model by specifying the class of functions f, for instance

$$\begin{aligned}
\mathcal{F}_0 &: \left\{ f(x) = \beta_0 \right\} \\
\mathcal{F}_1 &: \left\{ f(x) = \beta_0 + \beta_1 x \right\} \\
\mathcal{F}_{\sin} &: \left\{ f(x) = \sin(\omega x) \right\} \\
\mathcal{F}_{\exp} &: \left\{ f(x) = e^{-i\omega x} \right\}
\end{aligned}$$

- A model is characterized by its complexity and size
- We propose a class of model, among which we would like to find "the best"

Loss function and minimization criterion

- Define a loss function that scores the quality of fit of the model

$$\ell\big(\mathbf{Y}, f(\mathbf{X})\big)$$

- The loss function depends on the statistical model and on the objectives

$$\begin{aligned}\ell\big(\mathbf{Y}, f(\mathbf{X})\big) &= \frac{1}{n}\sum_{i=1}^{n} \|Y_i - f(X_i)\|_2^2, \quad \text{least squares} \\ &= \frac{1}{n}\sum_{i=1}^{n} \max\Big(0, 1 - Y_i f(X_i)\Big) \quad \text{hinge}\end{aligned}$$

- The estimator $\widehat{f}$ of a model f is a minimizer of the loss:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \big\{\ell\big(\mathbf{Y}, f(\mathbf{X})\big)\big\}$$

The train set

- Training set : the response and the features are jointly observed

$$\mathcal{D} = \big\{(Y_i, X_i), i = 1, \ldots, n\big\}$$

- An estimator is a function of the train set

$$\widehat{f}_{\mathcal{D}} = \underset{f \in \mathcal{F}, (\mathbf{Y}, \mathbf{X}) \in \mathcal{D}}{\arg\min} \big\{\ell\big(\mathbf{Y}, f(\mathbf{X})\big)\big\}$$

- Statistical properties of an estimator:

$$\mathbb{E}_{\mathbf{Y},\mathbf{X}}\Big(\widehat{f}_{\mathcal{D}}\Big) = \int xy \times \widehat{f}_{\mathcal{D}}(x, y) d\mathbb{P}(\mathbf{Y}, \mathbf{X})$$

Properties of an estimator

- Let us suppose that there is a true model: f^*
- The bias:

$$\text{Bias}(\widehat{f}) = \mathbb{E}_{\mathbf{Y},\mathbf{X}}\left(f^*(\mathbf{X}) - \widehat{f}(\mathbf{X})\right)$$

- The Variance:

$$\mathbb{V}(\widehat{f}) = \mathbb{V}_{\mathbf{Y},\mathbf{X}}\left(\widehat{f}(\mathbf{X})\right)$$

- The Mean Square error:

$$\text{MSE}(\widehat{f}) = \mathbb{E}_{\mathbf{Y},\mathbf{X}}\left(f^*(\mathbf{X}) - \widehat{f}(\mathbf{X})\right)^2 = \text{Bias}(\widehat{f})^2 + \mathbb{V}(\widehat{f})$$

- The prediction errors:

$$\text{Pred}(\widehat{f}) = \mathbb{E}_{\mathbf{Y},\mathbf{X}}\left(\mathbf{Y} - \widehat{f}(\mathbf{X})\right)^2$$

Properties of the LS estimator in linear regression

- The bias:

$$\text{Bias}(\widehat{\boldsymbol{\beta}}) = 0$$

- The Variance:

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- The Mean Square error:

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- The prediction errors:

$$\text{Pred}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \frac{p}{n} + \sigma^2$$

Risk of an estimator

- Central concept 1: the model collection does not necessarily catch the true model :

$$f^* \notin \mathcal{F}$$

- The oracle is the best possible estimator if all information was available
- From Vapnik and Chervonenskys ($\sim$1960s) the risk of an estimator

$$R(\widehat{f}) = \mathbb{E}\|f^* - \widehat{f}\|^2 = \mathbb{E}\|f^* - f_{\text{oracle}} + f_{\text{oracle}} - \widehat{f}\|^2$$

- The Approximation error

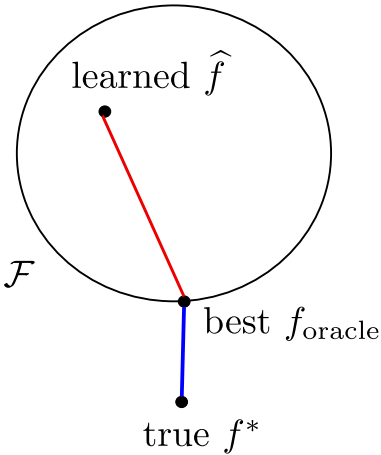
$$\mathbb{E}\|f^* - f_{\text{oracle}}\|^2$$

- Estimation error

$$\mathbb{E}\|f_{\text{oracle}} - \widehat{f}\|^2$$

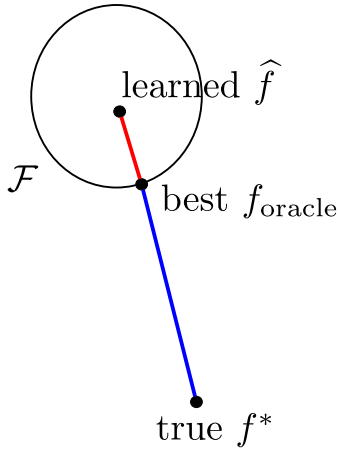
- How to control the estimation error ? How to compute the approximation error ?

High Complexity Model



Approx Err < Estimation Err

Low Complexity Model



Approx Err > Estimation Err

Outline

1. The multi-loci model, algebra of linear models
2. Increasing the dimension of linear models
3. The bias-variance trade-off
4. **Generalization error and cross validation**
5. Regularization and Penalization
6. Feature selection

Need to assess the generalization error

- The estimator has been trained on the training set $\mathcal{D}$:

$$\widehat{f}_{\mathcal{D}} = \underset{f \in \mathcal{F}, (\mathbf{Y}, \mathbf{X}) \in \mathcal{D}}{\arg\min} \left\{ \ell\big(\mathbf{Y}, f(\mathbf{X})\big) \right\}$$

- The estimator is expected to have good properties on the train set $\mathcal{D}$: the learning error

$$\text{Pred}\Big(\widehat{f}(\mathbf{X}_{\mathcal{D}})\Big) = \frac{1}{n_{\text{train}}} \left\| \mathbf{Y}_{\mathcal{D}} - \widehat{f}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}) \right\|^2 = \frac{1}{n_{\text{train}}} \left\| \mathbf{Y}_{\mathcal{D}} - \mathbf{X}_{\mathcal{D}} \widehat{\boldsymbol{\beta}}_{\mathcal{D}} \right\|^2$$

- What would be its performance on another set ? The test set $\mathcal{T}$

$$\text{Pred}\Big(\widehat{f}(\mathbf{X}_{\mathcal{T}})\Big) = \frac{1}{n_{\text{test}}} \left\| \mathbf{Y}_{\mathcal{T}} - \widehat{f}_{\mathcal{D}}(\mathbf{X}_{\mathcal{T}}) \right\|^2 = \frac{1}{n_{\text{test}}} \left\| \mathbf{Y}_{\mathcal{T}} - \mathbf{X}_{\mathcal{T}} \widehat{\boldsymbol{\beta}}_{\mathcal{D}} \right\|^2$$

- The Learning Error can dramatically underestimate the Generalization Error

Cross-validation : K-fold scheme

- Split the data into K partitions $\mathcal{I}_0 = \{1, \ldots, n\}$

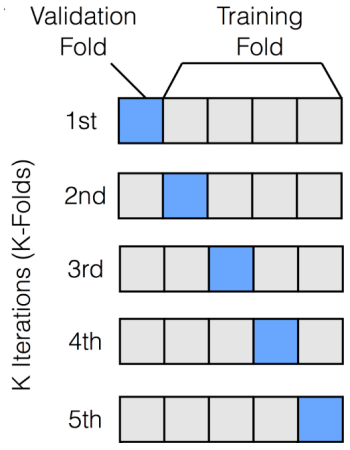
$$\mathcal{I}_0 = \bigcup_k \mathcal{I}_k, \quad |\mathcal{I}_k| \sim K/n$$

- For each fold, define a train and test

$$\mathcal{D}_k = \mathcal{I}_k, \quad \mathcal{T}_k = \mathcal{I}_0 \setminus \mathcal{D}_k$$

- Train the model on the train set

$$\widehat{f}_{\mathcal{D}_k} = \arg\min_{\mathcal{D}_k} \left\{ \ell\big(\mathbf{Y}, f(\mathbf{X})\big) \right\}$$



Cross-validation : K-fold scheme

- Estimate the generalization error on the test set

$$\widehat{\mathrm{Pred}}_k(\widehat{f}) = \frac{1}{|\mathcal{T}_k|} \sum_{i \in \mathcal{T}_k} \ell\big(Y_i, \widehat{f}_{\mathcal{D}_k}(X_i)\big)$$

- Average all prediction errors on all folds

$$\widehat{\mathrm{Pred}}(\widehat{f}) = \frac{1}{K} \sum_{k=1}^{K} \widehat{\mathrm{Pred}}_k(\widehat{f})$$

- This approach learns the model K times
- When $K = n$ this is called cross-validation
- Choose $K \sim 5 - 10$.

COMMENTARY

Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification

Richard Simon, Michael D. Radmacher, Kevin Dobbin, Lisa M. McShane

DNA microarrays have made it possible to estimate the level of expression of thousands of genes for a sample of cells. Although biomedical investigators have been quick to adopt this powerful new research tool, accurate analysis and interpretation of the data have provided unique challenges. Indeed, many investigators are not experienced in the analytical steps needed to convert tens of thousands of noisy data points into reliable and interpretable biologic information. Although some investigators recognize the importance of collaborating with experienced biostatisticians to analyze microarray data, the number and availability of experienced biostatisticians is inadequate. Consequently, investigators are using available software to analyze their data, many seemingly without knowledge of potential pitfalls. Because of serious problems associated with the analysis and reporting of some DNA microarray studies, there is great interest in guidance on valid and effective methods for analysis of DNA microarray data.

however, the emphasis is on developing a gene expression-based multivariate function (referred to as the predictor) that accurately predicts the class membership of a new sample on the basis of the expression levels of key genes. Such predictors can be used for many types of clinical management decisions, including risk assessment, diagnostic testing, prognostic stratification, and treatment selection. Many studies include both class comparison and class prediction objectives.

Class discovery is fundamentally different from class comparison or class prediction in that no classes are predefined. Usually the purpose of class discovery in cancer studies is to determine whether discrete subsets of a disease entity can be defined on the basis of gene expression profiles. This purpose is different from determining whether the gene expression profiles correlate with some already known diagnostic classification. Examples of class discovery are the studies by Bittner et al. *(3)* that examined gene expression profiles for advanced melanomas and by Alizadeh et al. *(4)* that examined the gene expression

Jacob and Speed *Genome Biology* (2018) 19:97
https://doi.org/10.1186/s13059-018-1481-6

Genome Biology

CORRESPONDENCE Open Access

The healthy ageing gene expression signature for Alzheimer's disease diagnosis: a random sampling perspective

Laurent Jacob[1*] and Terence P. Speed[2]

Abstract

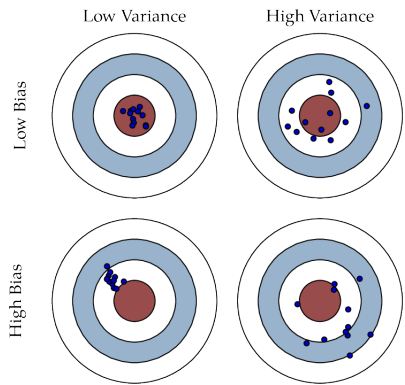
In a recent publication, Sood et al. (Genome Biol 16:185, 2015) presented a set of 150 probe sets that could be used in the diagnosis of Alzheimer's disease (AD) based on gene expression. We reproduce some of their experiments and show that their signature is indeed able to discriminate between AD and control patients using blood gene expression in two cohorts. We also show that its performance does not stand out compared to randomly sampled sets of 150 probe sets from the same array.

Outline

1. The multi-loci model, algebra of linear models
2. Increasing the dimension of linear models
3. The bias-variance trade-off
4. Generalization error and cross validation
5. **Regularization and Penalization**
6. Feature selection

Strategy

- The model with the highest number of parameter has low bias and high variance
- Prevent overfitting
- Improve interpretability
- Improve prediction accuracy



Penalized Empirical Risk

- The empirical risk minimization ensures low approximation error (low bias)

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell\big(Y_i, f(X_i)\big)$$

- We want to control the estimator errors (variance) by the complexity of the model
- Minimize the penalized empirical risk

$$\widehat{f}_\lambda = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i, f(X_i)\big) + \lambda \, \text{pen}(\mathcal{F})$$

- Accept some bias provided the variance is reduced: the shrinkage strategy

Preventing degeneracy with the Ridge penalty in Regression

- Come back to the linear regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

- Consider the penalized risk

$$R_\lambda(\mathbf{Y}; \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

- The Ridge estimator is

$$\widehat{\boldsymbol{\beta}}_\lambda^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$$

Intuition behind the ridge strategy

- When features are highly correlated and/or p increases, $\mathbf{X}'\mathbf{X}$ becomes degenerate
- The Ridge penalty regularized the empirical covariance matrix of features

$$\mathbf{S}_\lambda = \mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$$

- The LS estimator is such that

$$\widehat{\boldsymbol{\beta}}^{ridge} = \left(\mathbf{S}_\lambda^{-1}\mathbf{X}'\mathbf{X}\right)\widehat{\boldsymbol{\beta}}^{ls} = \left(\mathbf{I} - \lambda\mathbf{S}_\lambda^{-1}\right)\widehat{\boldsymbol{\beta}}^{LS}$$

- if $\lambda = 0$ we get the OLS.
- if features are independent $\mathbf{X}'\mathbf{X} = \mathbf{I}$

$$\widehat{\boldsymbol{\beta}}_\lambda^{ridge} = \frac{1}{1+\lambda}\widehat{\boldsymbol{\beta}}^{LS}$$

Intuition behind the ridge strategy

- The penalty introduces some bias
- The penalty reduces the variance:

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}^{ridge}) = \sigma^2 \mathbf{S}_\lambda^{-1} \times \mathbf{S} \times \mathbf{S}_\lambda^{-1} \leq \mathbb{V}(\widehat{\boldsymbol{\beta}}^{LS})$$

- When $\lambda \to 0$ there is no biais but the variance increases
- When $\lambda \to \infty$ there is high biais but the variance decreases
- A trade-off should be made by calibrating λ.
- Dependent variables are grouped
- Does not perform selection, appropriate for prediction (low interpretability)

Outline

1. The multi-loci model, algebra of linear models
2. Increasing the dimension of linear models
3. The bias-variance trade-off
4. Generalization error and cross validation
5. Regularization and Penalization
6. **Feature selection**

Feature selection

- This has been one of the hot topic in machine learning for years
- We can suppose that among p features, only s^* impact the response

$$\{1, \dots, p\} = \mathcal{S}_0 \cup \mathcal{S}_1$$

$$\mathcal{S}_0 : \{j,\, \beta_j^* = 0\}, \quad \mathcal{S}_1 : \{j,\, \beta_j^* \neq 0\}$$

- Suppose that the underlying efficient model would be

$$\mathbb{E}(\mathbf{Y} \mid \mathbf{X}) \simeq \mathbf{X}_{\mathcal{S}_1} \boldsymbol{\beta}_{\mathcal{S}_1} = \sum_{j \in \mathcal{S}_1} X_{ij} \beta_j$$

- Increase prediction accuracy, improve interpretability $s^* < p$.

Optimization strategies

- Many options: exhaustive, iterative approach (stepwise), forward-backward selection, stepwise
- Need an algorithm to find the subset and a model selection criterion to select the number of features
- The old ways : find the highest correlated feature

$$\begin{aligned} \widehat{j_1} &= \arg\max_j \text{corr}(\mathbf{Y}, \mathbf{x}^j) \\ \widehat{j_2} &= \arg\max_j \text{corr}(\mathbf{Y}, \mathbf{x}^j \mid \mathbf{x}^{\widehat{j_1}}) \end{aligned}$$

- Works if p is small, and cross correlations not too high **S**

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie

Stanford University, USA

[Received December 2003. Final revision September 2004]

Summary. We propose the elastic net, a new regularization and variable selection method. Real world data and a simulation study show that the elastic net often outperforms the lasso, while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the lasso is not a very satisfactory variable selection method in the $p \gg n$ case. An algorithm called LARS-EN is proposed for computing elastic net regularization paths efficiently, much like algorithm LARS does for the lasso.

Keywords: Grouping effect; LARS algorithm; Lasso; Penalization; $p \gg n$ problem; Variable selection

Selecting with the LASSO

- Least Absolute Shrinkage and Selection Operator
- Consider the penalized risk

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

- The solution is not explicit in the general case : need optimization techniques
- Different optimization strategies to find the best solution
- Has now become a standard in machine learning

What happens if features are independent ?

- The orthogonal case : $\mathbf{S} = \mathbf{X}'\mathbf{X} = \mathbf{I}$ the LASSO has an explicit solution

$$\widehat{\beta}_j^{lasso} = \max\left(0, 1 - \frac{\lambda}{|\widehat{\beta}_j^{LS}|}\right) \times \widehat{\beta}_j^{LS}$$

- This is called a thresholding estimator
- In the general case $\mathbf{S} \neq \mathbf{I}$, the LASSO thresholds features while accounting for their correlations

calibration of the hyperparameter

- Penalized estimators depend on the hyperparameter λ : should be tuned!
- Grid Search: compute $\widehat{\boldsymbol{\beta}}_\lambda$ for $\lambda \in [\lambda_{\min}, \lambda_{\max}]$
- Select the best of the best estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\lambda}}$
- A training data set is a dataset used during the learning process and is used to fit the parameters:
- A validation data set is a dataset used to tune the hyperparameters of predictor.
- A test set is an additional dataset used only to assess the performance (i.e. generalization) of a fully specified predictor.
- In all cases, after tuning the parameters and selecting a model, you should fit the chosen model on the complete dataset (not only on the train dataset) to define the final predictor.

Cross Validation for calibration [?]

Figure 3 | *K*-fold cross-validation involves splitting the data set into *K* subsets and doing multiple iterations of training and evaluation. The metric (for example, F_1 score) from all iterations is averaged. (**a**) A strategy with $K = 5$ without model selection. Training sets and test sets are used to derive prediction statistics. (**b**) Nested *K*-fold cross-validation with model selection. This strategy uses a validation set for model selection using the strategy of **a**. The best model is then tested on the separate test set. Gray bars indicate samples not used at the represented stage.

References

[1] J. Lever, M. Krzywinski, and Altman N. Model selection and overfitting. *Nat Methods*, 13:703–704, 2016.

[2] T. A. Stamey, J. N. Kabalin, J. E. McNeal, I. M. Johnstone, F. Freiha, E. A. Redwine, and N. Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *J Urol*, 141(5):1076–1083, May 1989.