Christophe Lalanne
www.aliquote.org

Avril 2012

# Synopsis

apprentissage statistique • sélection
de modèles • régularisation • approches
multivariées et sélection de variables • illustration

"To paraphrase provocatively, 'machine learning is statistics minus any checking of models and assumptions'. — Brian D. Ripley, useR! 2004, Vienna (May 2004)"

# Éléments de contexte

Breiman, 2001a : Statistical Modeling: The Two Cultures.

- "What's the model for the data?"
- "It is a strange phenomenon–once a model is made, then it becomes truth and the conclusions from it are infallible."
- "The whole area of guided regression is fraught with intellectual, statistical, computational, and subject matter difficulties."

La question qui se pose n'est pas l'usage de modèles dirigés par les données, mais bien du choix d'un modèle génératif, $P(x,y)$, ou discriminant, $P(y|x)$ (Ng et Jordan, 2001). Par ailleurs, la recherche d'un modèle prédictif *vs.* explicatif suggère le recours à des approches différentes.

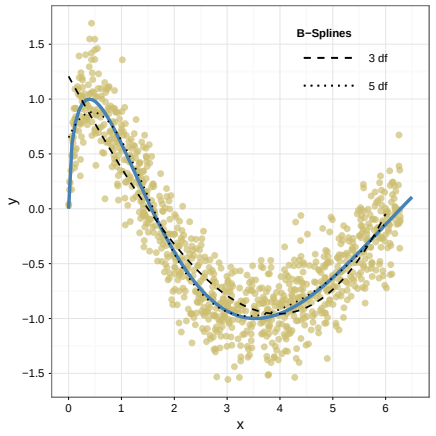
# Du problème du choix des paramètres

Modèle génératif :

```
f <- function(x) sin(sqrt(2*pi*x))
n <- 1000
x <- runif(n, 0, 2*pi)
y <- f(x) + rnorm(n, 0, 0.25)
```

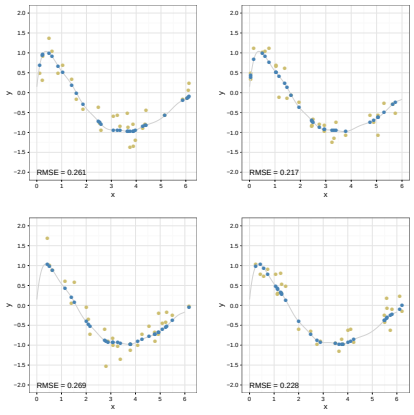
Qualité du modèle (N=1000 ou 30)

| Modèle | $R^2_a$ | AIC | DF |
|---|---|---|---|
| BS-3 | 0.8103 | 399.2 | 996 |
| BS-5 | 0.8562 | 124.2 | 994 |
| BS-10 | 0.8624 | 85.2 | 989 |
| BS-3 | 0.8377 | 20.6 | 26 |
| BS-5 | 0.8620 | 17.3 | 24 |
| BS-10 | 0.8716 | 18.2 | 19 |

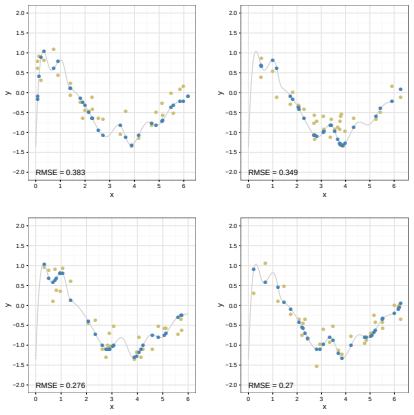


# Illustration

BS-15 estimé sur N=100, 30 nouvelles observations.




BS-15 estimé sur N=30, 30 nouvelles observations.

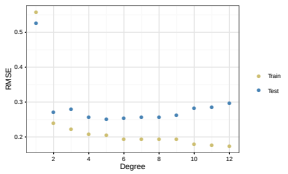



# Compromis biais/variance

Considérons un modèle polynomial d'ordre $k$, $f(x) = \beta_0 + \sum_{j=1}^{k} \beta_j x^j$.

Problématique de sélection de modèle :

- Large biais quand $k$ petit, large variance quand $k$ grand.
- Comment choisir $k$ ?




$$\mathbb{E}_X\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \underbrace{\left(y_0 - F(x_0)\right)^2}_{bruit} + \underbrace{\left(F(x_0) - \bar{f}(x_0)\right)^2}_{biais^2} + \underbrace{\mathbb{E}_X\left[\left(\hat{f}(x_0) - \bar{f}(x_0)\right)^2\right]}_{variance}$$

# Comment sélectionner/valider un modèle ?

Quel bon compromis pour éviter le sur-ajustement et contrôler la complexité du modèle ?

- Limiter le nombre de prédicteurs, $p < \frac{m}{15}$ (Harrell, 2001), ou maximiser $\log p(X \mid \hat{\beta}_{ML}) - \#\text{params}$ (AIC).
- Procédure de sélection automatique de variable ('stepwise').
- Validation croisée : 2 sous-échantillons, k-fold (avec ou sans répétition), bootstrap (Molinaro et al., 2005 ; Ambroise et McLachlan, 2002 ; Varma et Simon, 2006).
- Techniques de pénalisation, incluant la sélection automatique de variables.

Ces techniques se situent principalement dans un cadre de régression (G)LM.

# Approche de régression par régularisation

L'estimateur par MCO $(X'X)^{-1}X'Y$ minimise la SSR ($p \ll n$). Lorsque $p > n$, $X$ n'est pas de rang plein, et il n'y a plus unicité des solutions MCO. Pour la prédiction on peut chercher à minimiser le risque $\mathbb{E}(\mathbb{E}(y \mid x) - x'\beta)^2$.

Autre classe d'estimateurs : Ridge (Hoerl et Kennard, 1970), Bridge (Fu, 1998), Lasso (Tibshirani, 1996), Elasticnet (Zou et Hastie, 2005), SCAD (Fan et Li, 2001), Dantzig selector (Candes et Tao, 2007), Generalized Path Seeker (Friedman, 2008).

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \|Y - X\beta\|_2^2 + \underbrace{\lambda_1 \|\beta\|_1}_{\text{lasso}} + \underbrace{\lambda_2 \|\beta\|_2^2}_{\text{ridge}}$$

Une pénalisation avec une norme $L_0$, $\lambda \sum_j I(|\beta_j| \neq 0)$, revient aux méthodes AIC/BIC.

# Sélection de variables classique

Soit des données simulées selon le modèle $y = \beta_0 + \sum_{i=1}^{10} \beta_i x_i + \varepsilon$,

```
n <- 50
X <- replicate(10, rnorm(n))
y <- 1.1*X[,1] + 0.8*X[,2] - 0.7*X[,5] + 1.4*X[,6] + rnorm(n)
```

Typiquement, une procédure de sélection pas-à-pas (critère AIC), en considérant que notre modèle de base est `y ~ 0 + x1 + x2 +x3 + x4`, donnerait les résultats suivants :

```
  Step Df  Deviance Resid. Df Resid. Dev       AIC
1      NA        NA         46   96.46049 40.855522
2 + x6 -1 52.533786         45   43.92670  3.524967
3 + x5 -1  9.521129         44   34.40557 -6.690225
```

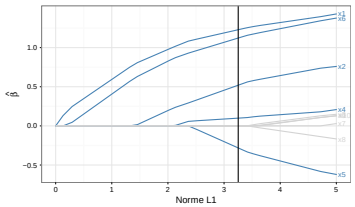
```
Residual standard error: 0.8843 on 44 degrees of freedom
Multiple R-squared: 0.8461, Adjusted R-squared: 0.8251
F-statistic: 40.32 on 6 and 44 DF,  p-value: 2.638e-16
```

| Term | $\hat{\beta}$ | SE | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| x1 | 1.3643 | 0.1361 | 10.02 | 0.0000 |
| x2 | 0.7970 | 0.1508 | 5.29 | 0.0000 |
| x3 | 0.1857 | 0.1280 | 1.45 | 0.1540 |
| x4 | 0.1981 | 0.1468 | 1.35 | 0.1842 |
| x6 | 1.2988 | 0.1483 | 8.76 | 0.0000 |
| x5 | -0.5881 | 0.1686 | -3.49 | 0.0011 |

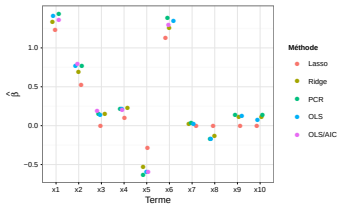
# Sélection de variables par régularisation

Optimisation directe d'un critère basé sur la vraisemblance (pénalisée), par régression *lasso* (Friedman et al., 2010).

| Méthode | RMSE (IQR) | $R^2$ (IQR) | Temps (s) |
|---|---|---|---|
| Lasso | 0.954 (0.76–1.13) | 0.862 (0.76–0.94) | 6.1 |
| Ridge | 0.928 (0.76–1.16) | 0.867 (0.76–0.93) | 26.6 |
| Elasticnet | 0.932 (0.73–1.14) | 0.866 (0.77–0.92) | 49.3 |




Solution *lasso* (optimisée par LOO-CV).




Comparaison des estimations.

# Colinéarité et grande dimension

Soit $Y$, réponse continue, et trois prédicteurs continus, $(X_1, X_2, X_3)$, mesurés sur un échantillon de taille $n = 80$, avec $(Y, X_1, X_2, X_3) \sim \mathcal{N}(0, \Sigma)$.
Considérons 40 prédicteurs additionnels tirés indépendemment dans $\mathcal{N}(0,1)$, de sorte qu'une estimation par simple MCO semble délicate.
Par ailleurs, si

$$\Sigma = \begin{pmatrix} 1 & -.5 & -.5 & 0 \\ -.5 & 1 & .5 & -.5 \\ -.5 & .5 & 1 & -.5 \\ 0 & -.5 & -.5 & 1 \end{pmatrix}$$

$X_3$ n'est pas corrélé à $Y$ mais sa corrélation partielle avec $Y$ n'est pas nulle.

Question : Est-il possible de recouvrer les prédicteurs 'intéressants' ?

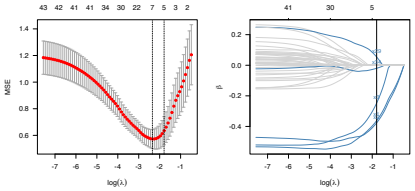
Filtrage univarié (test de corrélation) :
$X_3$ serait exclu de la liste des prédicteurs candidats (Paul et al., 2008).

| Term | $\log_{10}$(p) | Bonferroni | FDR |
|---|---|---|---|
| x1 | -7.2547459 | -5.621277 | -5.90614479 |
| x2 | -7.2385832 | -5.605115 | -5.90614479 |
| x3 | -0.6746860 | 0.000000 | -0.11700126 |
| x4 | -0.4135326 | 0.000000 | -0.11700126 |
| x4 | -0.1717955 | 0.000000 | -0.05684103 |

Approche par régularisation $L_1 L_2$, avec optimisation sur critère RMSE.

| Term | $\hat{\beta}_{OLS}$ (SE) | $\hat{\beta}_{enet}$ |
|---|---|---|
| x1 | -0.547 (0.083) | -0.362 |
| x2 | -0.548 (0.093) | -0.377 |
| x3 | -0.541 (0.093) | -0.248 |
| x29 | 0.216 (0.088) | 0.054 |
| x34 | -0.061 (0.071) | -0.020 |

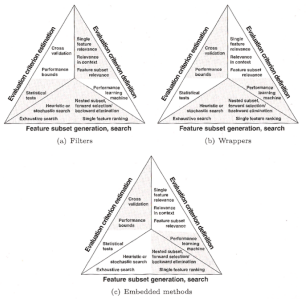
10-fold




# Autres approches pour la sélection de variables

- Méthodes de filtrage : généralement univariées, sélection de variables indépendamment du classifieur, extensions bayésiennes (Bo et Jonassen, 2002 ; Long et al., 2001).
- Méthodes d'ensemble ('wrapper') : qualité de la classification, importance des prédicteurs, impossible d'enrichir la structure des classifieurs (Lal et al., 2006).
- Méthodes intégrées ou enchâssées ('embedded') : processus de sélection de variables est intégré à l'algorithme d'apprentissage, moins exigeantes en termes de calcul (Lal et al., 2006).

Tiré de : Guyon et al., 2006




Ne pas négliger l'importance de la validation croisée (Schwender et al., 2008 ; Ambroise et McLachlan, 2002).

# Ensemble learning

L'idée est de générer des ensembles de classifieurs variés et suffisamment précis.

On peut introduire de la variabilité en variant différents paramètres :

- varier le poids des observations (boosting/bagging, Breiman, 1996)
- varier les valeurs des observations (perturbation par ajout de bruit)
- considérer des sous-ensembles de variables (random forests)
- varier les paramètres du modèle
- varier le modèle utiliser (arbres, MARS, NNs, etc.)

Les estimations peuvent ensuite être combinées par pondération des estimations, par une méthode de vote (en classification), ou par partitionnement de l'espace de design (Seni et Elder, 2010 ; Hastie et al., 2009).

# Approche non-paramétrique par arbres de décision

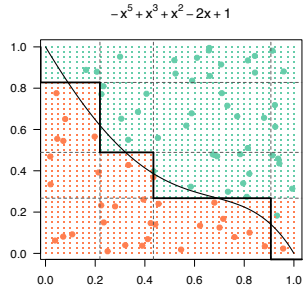
Dans le cas des structures de données irrégulières ($n \approx p$ ou $n \ll p$), les approches de filtrage univarié (tests t, régression) ou de réduction de dimension (PCA, SVD) ne prennent pas en compte la nature multivariée du problème.

Avantages des arbres de décision (CART, ID3, C4.5/J48, etc.) :

- fonctionnent avec des variables numériques ou qualitatives, avec ou sans valeurs manquantes, moins sensibles aux valeurs extrêmes,
- capturent les interactions, ignorent les prédicteurs de faible poids.

En revanche, ils sont instables, ne capturent pas bien les combinaisons linéaires de variables, et sont impactés par la colinéarité (*surrogates*).

Typiquement, il est nécessaire d'élaguer l'arbre de décision pour éviter le sur-ajustement : minimiser taille de l'arbre + minimiser fonction de coût.




```
1) root 80 19.3875000 0.58750000
  2) y< 0.2673223 24  0.9583333 0.04166667
    4) x< 0.9072912 22  0.0000000 0.00000000 *
    5) x>=0.9072912 2  0.5000000 0.50000000 *
  3) y>=0.2673223 56  8.2142860 0.82142860
    6) x< 0.2191463 10  1.6000000 0.20000000
     12) y< 0.8277749 8  0.0000000 0.00000000 *
     13) y>=0.8277749 2  0.0000000 1.00000000 *
    7) x>=0.2191463 46  1.9130430 0.95652170
     14) x< 0.4344256 9  1.5555560 0.77777780
       28) y< 0.4891134 2  0.0000000 0.00000000 *
       29) y>=0.4891134 7  0.0000000 1.00000000 *
     15) x>=0.4344256 37  0.0000000 1.00000000 *
```

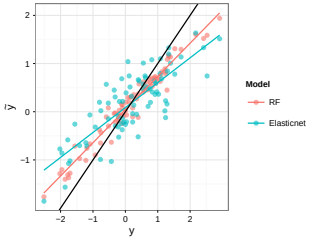
## Random Forests

Extension des CART incluant une double étape de randomization (variables et individus). Il n'y a pas de modèle sous-jacent : il s'agit d'un algorithme (Breiman, 2001b) :

- On spécifie le nombre de variables $p$ qui servira d'ensemble de prédicteurs parmi les $P$ variables de départ (généralement, $p \sim \sqrt{P}$).
- Chaque arbre (de profondeur maximale) est construit à partir d'un échantillon bootstrap des individus de l'ensemble d'apprentissage.
- À chaque nœud, $p$ variables sont sélectionnées aléatoirement parmi les $N$ variables, la division de l'arbre se faisant selon un critère de maximisation du gain d'information sur ces $n$ variables (Gini : $i(\Omega) = \sum_{r \neq s} p(r \mid \Omega) p(s \mid \Omega)$ ou $\frac{1}{n_\Omega} \sum_{i \in \Omega} (y_i - \bar{y})^2$).
- L'importance de chaque variable est évaluée par permutation.

## Random Forests et régression

Avec les données de régression 80 x 43, on peut comparer les variables 'sélectionnées' par régression $L_1 L_2$ et par RF :




| Méthode | RMSE (IQR) | $R^2$ (IQR) | Temps (s) |
|---|---|---|---|
| RF | 0.977 (0.87–0.97) | 0.295 (0.19–0.37) | 304.6 |
| Elasticnet | 0.830 (0.88–1.03) | 0.451 (0.36–0.54) | 58.1 |

| Terme | Inc MSE (%) | Inc MSE (%) |
|---|---|---|
| x2 | 14.39 (18.5) | 0.36 (10.8) |
| x1 | 12.15 (15.6) | 0.36 (10.5) |
| x34 | 5.19 (6.7) | 0.19 (5.7) |
| x3 | 2.68 (3.4) | 0.09 (2.7) |
| x16 | 2.20 (2.8) | 0.03 (1.0) |

Les mesures d'importance peuvent également être utilisées pour effectuer de la sélection automatique (descendante) de variables (Díaz-Uriarte et Alvarez de Andrés, 2006).

## Refining Developmental Coordination Disorder subtyping with multivariate statistical method

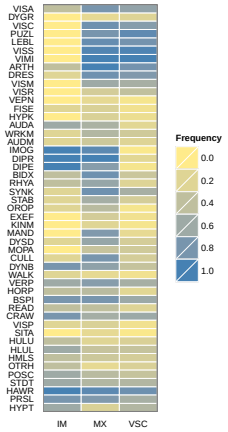
**Background:** With a large number of potentially relevant clinical indicators penalization and ensemble learning methods are thought to provide better predictive performance than usual linear predictors. However, little is known about how they perform in clinical studies where few cases are available. We used Random Forests and Partial Least Squares Discriminant Analysis to select the most salient impairments in Developmental Coordination Disorder (DCD) and assess patients similarity.
**Methods:** We considered a wide-range testing battery for various neuropsychological and visuo-motor impairments which aimed at characterizing subtypes of DCD in a sample of 63 children. Classifiers were optimized on a training sample, and they were used subsequently to rank the 49 items according to a permuted measure of variable importance. In addition, subtyping consistency was assessed with cluster analysis on the training sample. Clustering fitness and predictive accuracy were evaluated on the validation sample.
**Results:** Both classifiers yielded a relevant subset of items impairments that altogether accounted for a sharp discrimination between three DCD subtypes: ideomotor, visual-spatial and constructional, and mixt dyspraxia. The main impairments that were found to characterize the three subtypes were: digital perception, imitations of gestures, digital praxia, lego blocks, visual spatial structuration, visual motor integration, coordination between upper and lower limbs. Classification accuracy was above 90% for all classifiers, and clustering fitness was found to be satisfactory.
**Conclusion:** Random Forests and Partial Least Squares Discriminant Analysis are useful tools to extract salient features from a large pool of correlated binary predictors, but also provide a way to assess individuals proximities in a reduced factor space. Less than 15 neuro-visual, neuro-psychomotor and neuro-psychological tests might be required to provide a sensitive and specific diagnostic of DCD on this particular sample, and isolated markers might be used to refine our understanding of DCD in future studies.

Soumis à *BMC Medical Research Methodology*.

# Données sur l'ensemble de l'échantillon

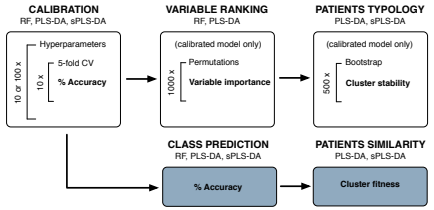
Fréquence d'échec par groupe clinique

| Item | IM | VSC | MX | FDR | Bonferroni |
|---|---|---|---|---|---|
| LEBL | 0.0 | 0.82 | 0.88 | 0.004895105 | 0.04895105 |
| PUZL | 0.0 | 0.91 | 0.80 | 0.004895105 | 0.04895105 |
| ARTH | 0.4 | 0.82 | 1.00 | 0.004895105 | 0.04895105 |
| DIPR | 1.0 | 0.18 | 1.00 | 0.004895105 | 0.04895105 |
| IMOG | 1.0 | 0.03 | 0.92 | 0.004895105 | 0.04895105 |
| DIPE | 1.0 | 0.06 | 0.72 | 0.004895105 | 0.04895105 |
| CULL | 0.2 | 0.27 | 0.80 | 0.004895105 | 0.04895105 |
| MAND | 0.0 | 0.15 | 0.76 | 0.004895105 | 0.04895105 |
| VIMI | 0.0 | 1.00 | 1.00 | 0.004895105 | 0.04895105 |
| VISS | 0.0 | 0.97 | 0.96 | 0.004895105 | 0.04895105 |
| BIDX | 0.4 | 0.33 | 0.84 | 0.008158508 | 0.09790210 |
| VISC | 0.0 | 0.73 | 0.84 | 0.008158508 | 0.09790210 |
| OROP | 0.2 | 0.06 | 0.44 | 0.017482517 | 0.24475524 |
| RHYA | 0.4 | 0.21 | 0.64 | 0.017482517 | 0.24475524 |
| SYNK | 0.2 | 0.55 | 0.84 | 0.029370629 | 0.44055944 |
| DYSD | 0.2 | 0.27 | 0.64 | 0.031674208 | 0.53846154 |
| DRES | 0.2 | 0.76 | 0.84 | 0.031674208 | 0.53846154 |
| AUDA | 0.6 | 0.18 | 0.52 | 0.043512044 | 0.78321678 |

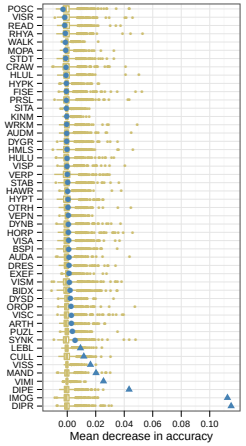



# Construction et évaluation des modèles

- 2/3 des individus pour l'apprentissage et 1/3 pour le test, stratifié sur le groupe clinique
- modèles : forêts aléatoires (RF) + analyse discriminante (s)PLS (avec ou sans Lasso)
- validation croisée des hyperparamètres (*mtry* et nombre de composantes) des modèles à partir du taux de classification correcte
- sélection de variables à partir de tests de permutation
- classification automatique optimisée par ré-échantillonnage
- taux de classification correcte, Sensibilité/Spécificité, et affinité par classe




# Variables isolées par RF et PLS




Variables sélectionnées par Lasso en sPLS-DA :

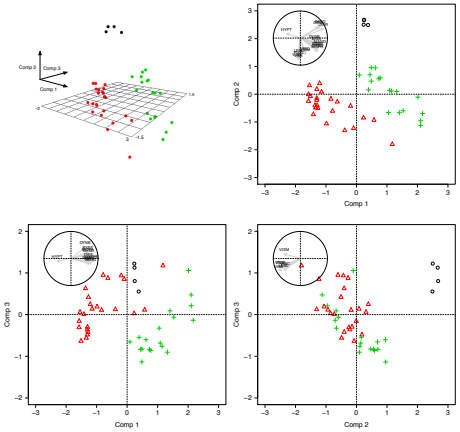
| Item | IM | VSC | MX |
|---|---|---|---|
| Imitation of gestures* | 100.00 | 88.406 | 88.406 |
| Visual spatial structuration* | 98.55 | 5.797 | 0.000 |
| Lego blocks* | 95.65 | 2.899 | 2.899 |
| Puzzles | 95.65 | 8.696 | 8.696 |
| Digital perception* | 95.65 | 66.667 | 66.667 |
| Digital praxia* | 91.30 | 0.000 | 91.304 |
| Visual spatial constructional | 86.96 | 8.696 | 8.696 |
| Coordination upper/lower limbs* | 81.16 | 59.420 | 59.420 |
| Synkinesis | 81.16 | 15.942 | 15.942 |
| Manual dexterity* | 81.16 | 63.768 | 63.768 |

* identique à RF

RF : *mtry* $= 12$, performance $92.4 \pm 5.5$ %.

sPLS : 2 composantes ($\eta = 0.7$), $94.2 \pm 7.6$ %.

# Profiles factoriels (PLS)



# Conclusions

- Les approches de régression pénalisée permettent de traiter des cas complexes et intègrent la sélection automatique de variables. En revanche, les estimateurs sont biaisés et il est difficile d'y associer des intervalles de confiance (Bunea et al., 2011 ; Kyung et al., 2010).
- Il est important de contrôler le risque de sur-ajustement, en particulier lorsque l'on couple des procédures de sélection de variables à des algorithmes de classification ou de régression. Une solution est d'utiliser des algorithmes qui intègrent la sélection ou le classement de variables.
- La comparaison des performances de différents classifieurs doit et peut être contrôlée (Hothorn et al., 2005 ; Jelizarow et al., 2010).

# Références


Ambroise, C. et McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566.

Bo, T. et Jonassen, I. (2002). New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 3(4):0017.1-0017.11.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26:123-140.

Breiman, L. (2001b). Random forests. *Machine Learning*, 45(1):5-32.

Breiman, L. (2001a). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

Bunea, F., She, Y., Ombao, H., Gongvatana, A. et Devlin, K. et al. (2011). Penalized least squares regression methods and applications to neuroimaging. *NeuroImage*, 55(4):1519–1527.

Candes, E. et Tao, T. (2007). The Dantzig selector: statistical estimation when $p$ is much larger than $n$. *The Annals of Statistics*, 35(6):2312–2351.

Díaz-Uriarte, R. et Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3.

Fan, J. et Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Friedman, J. (2008). Fast sparse regression and classification. Dans *In Proceedings of the 23rd International Workshop on Statistical Modelling*, pages 27–57.

Friedman, J., Hastie, T. et Tibsharini, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

Fu, W. (1998). Penalized regressions: The bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Guyon, I., Gunn, S., Nikravesh, M. et Zadeh, L. A., éditeurs (2006). *Feature Extraction: Foundations And Applications* Springer-Verlag.

Harrell, F. (2001). *Regression Modeling Strategies*. Springer.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, 2ème édition.

Hoerl, A. et Kennard, R. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hothorn, T., Leisch, F., Zeilis, A. et Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675-699.

Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K. et Boulesteix, A.-L. (2010). Over-optimism in bioinformatics: an illustration. *Bioinformatics*, 26(16):1990–1998.

Kyung, M., Gill, J., Ghosh, M. et Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–412.

Lal, T. N., Chapelle, O., Weston, J. et Elisseeff, A. (2006). Embedded methods. Dans Guyon, I., Gunn, S., Nikravesh, M. et Zadeh, L. A., éditeurs, *Feature Extraction: Foundations And Applications*, pages 137-162. Springer-Verlag.

Long, A., Mangalam, H., Chan, B., Tolleri, L. et Hatfield, G. et al. (2001). Improved statistical inference from dna microarray data using analysis of variance and a Bayesian statistical framework. *Journal of Biological Chemistry*, 276:19937-19944.

Molinaro, A. M., Simon, R. et Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, 21(15):3301-3307.

Ng, A. et Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Neural Information Processing Systems*, pages 841–848.

Paul, D., Bair, E., Hastie, T. et Tibshirani, R. (2008). Pre-conditioning for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 36:1595–1618.

Schwender, H., Ickstadt, K. et Rahnenführer (2008). Classification with high-dimensional genetic data: Assigning patients and genetic features to known classes. *Biometrical Journal*, 6:911–926.

Seni, G. et Elder, J. (2010). *Ensemble Methods in Data Mining*. Morgan & Claypool.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Varma, S. et Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(91).

Zou, H. et Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320.


Articles et code R disponibles à l'adresse suivante :
`http://aliquote.org/articles/slides/mva_clinres`.