Adaptive Handling of Dependence

in High-Dimensional Regression Modeling

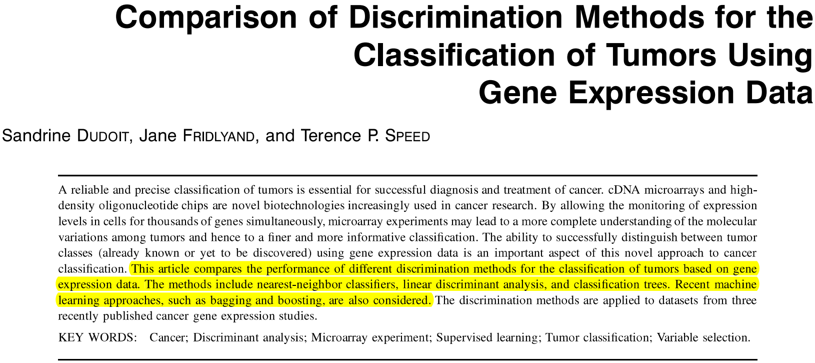
David Causeur
IRMAR, UMR 6625 CNRS
Agrocampus, Rennes, France

Joint work with Florian Hébert and Mathieu Emily

Outline of the talk

- Three short stories
    - In praise of laziness
    - Dependence is a blessing
    - The truth lies elsewhere
- An adaptive handling of dependence
    - The naïve option
    - A new class $\mathcal{L}$ of linear prediction scores
    - Optimal prediction within $\mathcal{L}$

In praise of laziness



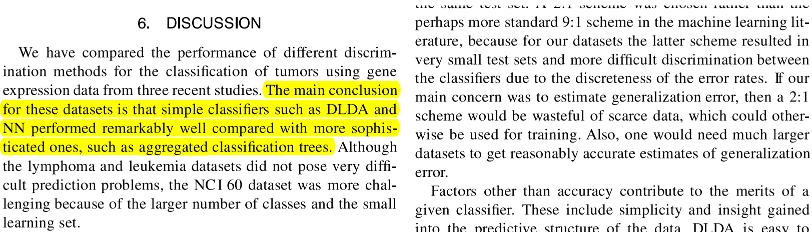
Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data

Sandrine Dudoit, Jane Fridlyand, and Terence P. Speed

A reliable and precise classification of tumors is essential for successful diagnosis and treatment of cancer. cDNA microarrays and high-density oligonucleotide chips are novel biotechnologies increasingly used in cancer research. By allowing the monitoring of expression levels in cells for thousands of genes simultaneously, microarray experiments may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more informative classification. The ability to successfully distinguish between tumor classes (already known or yet to be discovered) using gene expression data is an important aspect of this novel approach to cancer classification. This article compares the performance of different discrimination methods for the classification of tumors based on gene expression data. The methods include nearest-neighbor classifiers, linear discriminant analysis, and classification trees. Recent machine learning approaches, such as bagging and boosting, are also considered. The discrimination methods are applied to datasets from three recently published cancer gene expression studies.

KEY WORDS: Cancer; Discriminant analysis; Microarray experiment; Supervised learning; Tumor classification; Variable selection.

In praise of laziness



6. DISCUSSION

We have compared the performance of different discrimination methods for the classification of tumors using gene expression data from three recent studies. The main conclusion for these datasets is that simple classifiers such as DLDA and NN performed remarkably well compared with more sophisticated ones, such as aggregated classification trees. Although the lymphoma and leukemia datasets did not pose very difficult prediction problems, the NCI 60 dataset was more challenging because of the larger number of classes and the small learning set.

the same test set. A 2:1 scheme was chosen rather than the perhaps more standard 9:1 scheme in the machine learning literature, because for our datasets the latter scheme resulted in very small test sets and more difficult discrimination between the classifiers due to the discreteness of the error rates. If our main concern was to estimate generalization error, then a 2:1 scheme would be wasteful of scarce data, which could otherwise be used for training. Also, one would need much larger datasets to get reasonably accurate estimates of generalization error.

Factors other than accuracy contribute to the merits of a given classifier. These include simplicity and insight gained into the predictive structure of the data. DLDA is easy to

In praise of laziness

☺ *The lazy option can be the best* ☺

Ignoring dependence 1 - 0 Not ignoring dependence

Dependence is a blessing

Global testing *a.k.a* Signal detection

- p null hypotheses $H_0^{(j)}:\ \gamma_j = 0,\ j = 1, \ldots, p$.
- p pointwise test statistics $\boldsymbol{T} = (T_1, \ldots, T_p)$.
- Global testing is the test of $H_0:\ \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) = 0$
- by aggregating the pointwise $T_j,\ j = 1, \ldots, p$.

Functional ANOVA is a special case of global testing

Dependence is a blessing

The Higher Criticism [Donoho and Jin, 2004, 2008]

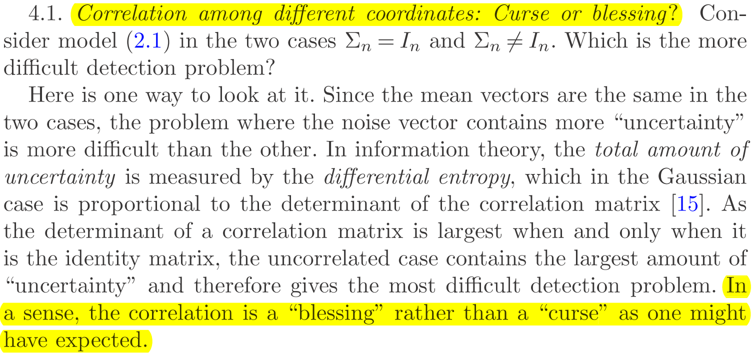
- The Rare-and-Weak paradigm : $\boldsymbol{T} \sim \mathcal{N}\left(\boldsymbol{\mu}; \boldsymbol{\Sigma}\right)$ with
  - A small fraction of non-zero coordinates in $\boldsymbol{\mu}$
  - Coordinates of $\boldsymbol{\mu}$ have small amplitudes
- The Higher-Criticism global test statistics :

$$\text{HC} = \max_{j : \frac{1}{n} \leq p_{(j)} \leq \frac{1}{2}} \sqrt{n} \frac{\frac{j}{n} - p_{(j)}}{\sqrt{p_{(j)}\left(1 - p_{(j)}\right)}}$$

- Reaches optimal (Chernoff) detection bounds when $\boldsymbol{\Sigma}$ is diagonal.

Dependence is a blessing

Under dependence : Hall and Jin (2010)

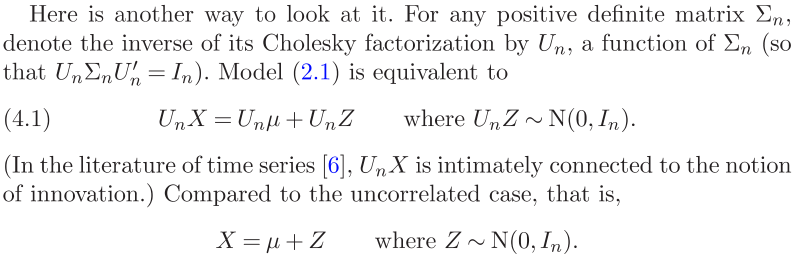


4.1. *Correlation among different coordinates: Curse or blessing?* Consider model (2.1) in the two cases $\Sigma_n = I_n$ and $\Sigma_n \neq I_n$. Which is the more difficult detection problem?

Here is one way to look at it. Since the mean vectors are the same in the two cases, the problem where the noise vector contains more "uncertainty" is more difficult than the other. In information theory, the *total amount of uncertainty* is measured by the *differential entropy*, which in the Gaussian case is proportional to the determinant of the correlation matrix [15]. As the determinant of a correlation matrix is largest when and only when it is the identity matrix, the uncorrelated case contains the largest amount of "uncertainty" and therefore gives the most difficult detection problem. In a sense, the correlation is a "blessing" rather than a "curse" as one might have expected.

Dependence is a blessing

Under dependence : Hall and Jin (2010)



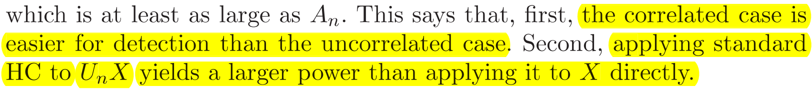
Here is another way to look at it. For any positive definite matrix Σ_n, denote the inverse of its Cholesky factorization by U_n, a function of Σ_n (so that $U_n \Sigma_n U_n' = I_n$). Model (2.1) is equivalent to

$$U_n X = U_n \mu + U_n Z \qquad \text{where } U_n Z \sim \mathrm{N}(0, I_n). \tag{4.1}$$

(In the literature of time series [6], $U_n X$ is intimately connected to the notion of innovation.) Compared to the uncorrelated case, that is,

$$X = \mu + Z \qquad \text{where } Z \sim \mathrm{N}(0, I_n).$$

...



which is at least as large as A_n. This says that, first, the correlated case is easier for detection than the uncorrelated case. Second, applying standard HC to $U_n X$ yields a larger power than applying it to X directly.

Dependence is a blessing

Whitening enhances detectability !

Ignoring dependence 1 - 1 Not ignoring dependence

The Phonological Neighborhood Density study

Phonological Neighborhood Density (PND) of a word : number of words that can be generated by replacing a phoneme with another phoneme in the same position.

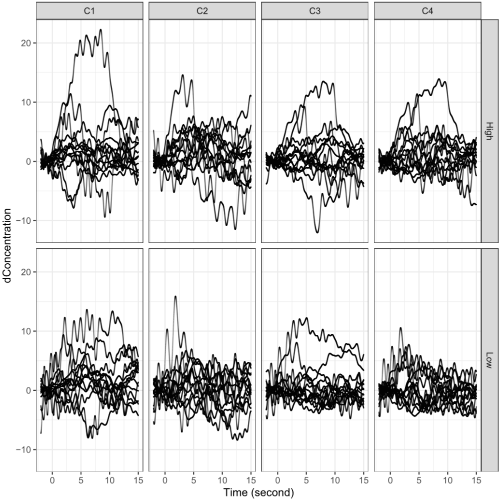
Examples : PROUD has a high PND, PROMPT has a low PND

Words with high PND : [(Chen *et al.*, 2011)]

- are recognized more slowly ;
- elicits greater changes in blood oxygenation in the left than in the right hemisphere of the brain.

The Phonological Neighborhood Density study

PND × channels hemodynamic curve data for 14 subjects.



The Phonological Neighborhood Density study

The linear function-to-scalar regression framework

- Hemodynamic response curve : $\boldsymbol{Y} = \big(Y(t_1), \ldots, Y(t_p)\big)'$
- Channel, Brain side, Subject effects : $\boldsymbol{x} = (x_1, \ldots, x_m)'$

$$\boldsymbol{Y} \;=\; \boldsymbol{\beta}\boldsymbol{x} + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\varepsilon} \sim \mathcal{N}(0; \boldsymbol{\Sigma})$$

ℓ_1-penalized deviance estimation

[(Rothman *et al.*, 2010)]

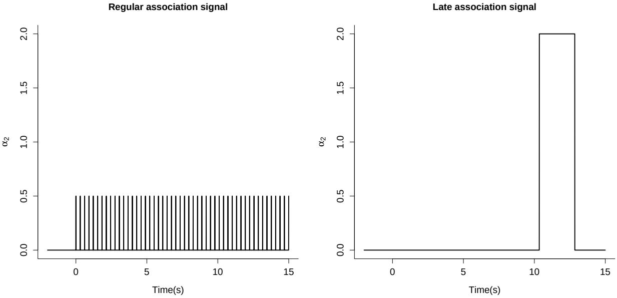
$$\mathcal{D}(\boldsymbol{\beta}; \boldsymbol{\Sigma}, \kappa) \;=\; n \log \det(\boldsymbol{\Sigma}) + \sum_{i=1}^{n} (\boldsymbol{Y}_i - \boldsymbol{\beta}\boldsymbol{x}_i)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\beta}\boldsymbol{x}_i) + \kappa ||\boldsymbol{\beta}||_1,$$

where $\kappa > 0$ is the penalty parameter.

The Phonological Neighborhood Density study

How does the choice of $\Omega = \Sigma^{-1}$ affect estimation ?

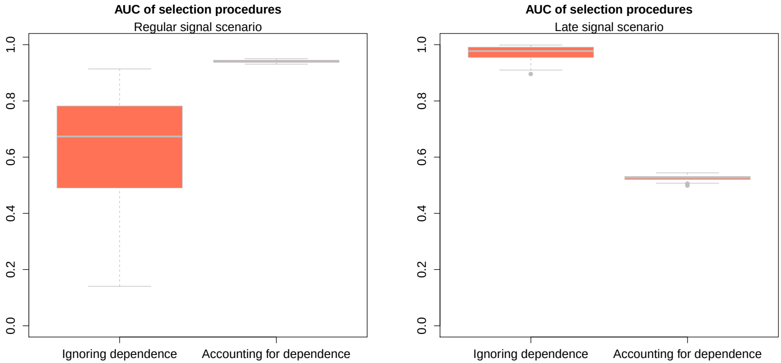
- Two options for High vs Low PND difference curve :



- Two options for $\mathcal{D}(\boldsymbol{\beta}; \boldsymbol{\Sigma}, \kappa)$:
    - A diagonal $\boldsymbol{\Sigma}$;
    - A close factor approximation of the sample estimate of $\boldsymbol{\Sigma}$.

The Phonological Neighborhood Density study

Focus on feature selection



The Phonological Neighborhood Density study

Ignoring dependence or not depends on the interplay of the patterns of dependence and association signal.

The good old normal linear regression setup

Linear regression settings

- Y a (numeric) response variable
- $\boldsymbol{X} = (X_1, \ldots, X_p)'$ a p-vector of explanatory variables

$$\begin{pmatrix} \boldsymbol{X} \\ Y \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix} ; \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\sigma}_{xy} \\ \boldsymbol{\sigma}'_{xy} & \sigma_y^2 \end{pmatrix} \right\}$$

Optimal linear prediction score (Oracle) :

$$\begin{aligned} L_{\text{opt}}(\boldsymbol{X}) &= \mu_y + (\boldsymbol{X} - \boldsymbol{\mu}_x)'\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\sigma}_{xy}, \\ &\equiv (\boldsymbol{X} - \boldsymbol{\mu}_x)'\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\sigma}_{xy} \end{aligned}$$

with a focus on $\text{R}^2 = \text{cor}^2(Y, L(\boldsymbol{X}))$.

The good old normal linear regression setup

Linear regression settings

- Y a (numeric) response variable
- $\boldsymbol{X} = (X_1, \ldots, X_p)'$ a p-vector of explanatory variables

$$\begin{pmatrix} \boldsymbol{X} \\ Y \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix} ; \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\sigma}_{xy} \\ \boldsymbol{\sigma}'_{xy} & \sigma^2_y \end{pmatrix} \right\}$$

Optimal linear prediction score (Oracle)

$$L_{\text{opt}}(\boldsymbol{X}) \equiv (\boldsymbol{X} - \boldsymbol{\mu}_x)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_{xy},$$

with conditional variance-covariance matrix of $\boldsymbol{X}$ given Y :

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_x - \frac{\boldsymbol{\sigma}_{xy}\boldsymbol{\sigma}'_{xy}}{\sigma^2_y}$$

The naïve linear prediction score

Optimal linear prediction score

$$\begin{aligned} L_{\text{opt}}(\boldsymbol{X}) &\equiv (\boldsymbol{X}-\boldsymbol{\mu}_x)'\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{R}^{-1}\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\sigma}_{xy}, \text{ where } \boldsymbol{R}=\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\Sigma\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}, \\ &\equiv (\boldsymbol{X}-\boldsymbol{\mu}_x)'\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{\lambda}}^{-1}\boldsymbol{U}'\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\sigma}_{xy}, \\ &\quad \text{where } \boldsymbol{R}=\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{\lambda}}\boldsymbol{U}' \text{ is the SVD of } \boldsymbol{R}, \\ &\equiv \boldsymbol{Z}'\boldsymbol{D}_{\boldsymbol{\lambda}}^{-1}\gamma \end{aligned}$$

where $\boldsymbol{Z}$ is a whitened version of $\boldsymbol{X}$: $\text{Var}(\boldsymbol{Z})=\boldsymbol{D}_{\boldsymbol{\lambda}}$.

Naïve linear prediction score

$$\begin{aligned} L_{\text{N}}(\boldsymbol{X}) &\equiv (\boldsymbol{X}-\boldsymbol{\mu}_x)'\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{I}_p\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\sigma}_{xy}, \\ &\equiv (\boldsymbol{X}-\boldsymbol{\mu}_x)'\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{U}\boldsymbol{U}'\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\sigma}_{xy}, \\ &\equiv \boldsymbol{Z}'\gamma \end{aligned}$$

Naïve vs optimal linear prediction score

Relative efficiency (over all γ)

Let $\boldsymbol{v}(\boldsymbol{\lambda})$ denote the eigenvector associated to the only positive eigenvalue of $\boldsymbol{\lambda}\boldsymbol{\lambda}^{-1\prime} + \boldsymbol{\lambda}^{-1}\boldsymbol{\lambda}'$ with $\boldsymbol{v}(\boldsymbol{\lambda})'\boldsymbol{v}(\boldsymbol{\lambda}) = 1$. Then,

$$\frac{f(\mathsf{R}^2_{\mathsf{opt}}) + 1}{f(\mathsf{R}^2_{\mathsf{opt}}) + g_{\max}(\boldsymbol{\lambda})} \leq \frac{\mathsf{R}^2_{\mathsf{N}}}{\mathsf{R}^2_{\mathsf{opt}}} \leq 1,$$

where $g_{\max}(\boldsymbol{\lambda}) = \boldsymbol{v}(\boldsymbol{\lambda})'\boldsymbol{\lambda}.\boldsymbol{v}(\boldsymbol{\lambda})'\boldsymbol{\lambda}^{-1}$ and $f(x) = x/(1-x)$.

Sharp bounds

- Worst case : If $\gamma = \boldsymbol{v}^{1/2}(\lambda)$, then $\mathsf{R}^2_{\mathsf{N}}$ reaches its lower limit.
- Best case : for any vector γ with only one nonzero coordinate, $\mathsf{R}^2_{\mathsf{N}}$ reaches its upper limit.

Naïve vs optimal linear prediction score

Comparison study (based on data-driven simulations)

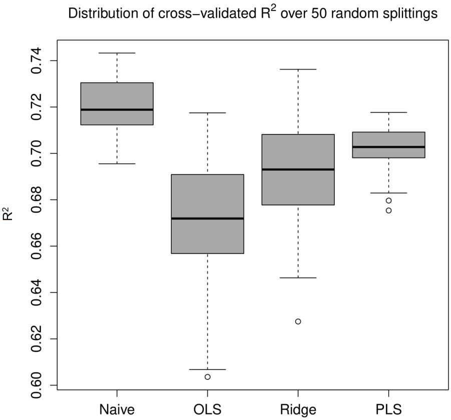
- $(\boldsymbol{\Sigma}, \boldsymbol{\sigma}_{xy}, \sigma_y^2)$ estimated using a public gene expression dataset [lu *et al.*, 2004] with $n = 30$ and $p = 403$
- Two scenarios for γ : worst and best case

		Naive	OLS	Ridge	PLS
Worst case	$n = 1000$	0.23 [0.09, 0.53]	0.79 [0.77, 0.81]	0.79 [0.77, 0.81]	0.79 [0.76, 0.81]
	$n = 30$	0.13 [0,0.44]	0.22 [0, 0.53]	0.55 [0.31, 0.69]	0.38 [0.02, 0.66]
Best case	$n = 1000$	0.80 [0.78, 0.82]	0.79 [0.77,0.82]	0.80 [0.78, 0.82]	0.80 [0.78, 0.82]
	$n = 30$	0.80 [0.78, 0.82]	0.23 [0, 0.53]	0.78 [0.64, 0.82]	0.80 [0.78, 0.82]

Naïve vs optimal linear prediction score

Comparison study (with the real Y, age of a subject)

Distribution of CV'd R^2 over 50 random splittings



A new class $\mathcal{L}$ of linear prediction scores

It all started with the following observations :

$$\begin{aligned} L_{\text{OLS}}(\hat{\mathbf{Z}}) &\equiv \hat{\mathbf{Z}}' D^{-}_{\hat{\boldsymbol{\lambda}}} \hat{\gamma} = \hat{\boldsymbol{\lambda}}^{-1'} \xi(\hat{\mathbf{Z}}), \\ L_{\text{N}}(\hat{\mathbf{Z}}) &\equiv \hat{\mathbf{Z}}' \hat{\gamma} = \mathbf{1}' \xi(\hat{\mathbf{Z}}) \end{aligned}$$

where $\xi(\hat{Z}) = \hat{Z} \odot \hat{\gamma}$ and $\odot$ stands for the term-by-term product.

A wide scope of dependence handling strategies is covered by :

$$\mathcal{L} = \left\{ L_{\boldsymbol{h}}(\hat{\mathbf{Z}}) = \boldsymbol{h}' \xi(\hat{Z}),\ \boldsymbol{h} = (h_1, \ldots, h_p)',\ \text{with } \boldsymbol{h}'\boldsymbol{h} = 1 \right\}.$$

Note : $\mathcal{L} \subset \{\ell' \boldsymbol{X}, \ell \in \mathbb{R}^p\}$.

A new class $\mathcal{L}$ of linear prediction scores

$\mathcal{L}$ contains Ridge prediction scores

$$L_{\text{Ridge}}(\hat{\mathbf{Z}}, \kappa) \quad \equiv \quad \boldsymbol{h}'_\kappa \xi(\hat{Z}).$$

with

- $\lim_{\kappa \to +\infty} \boldsymbol{h}_\kappa = (1/\sqrt{p})\mathbf{1}_p$ ... leading to $L_{\text{N}}(\hat{\mathbf{Z}})$
- $\lim_{\kappa \to 0} \boldsymbol{h}_\kappa = \hat{\boldsymbol{\lambda}}^{-1} / \sqrt{\hat{\boldsymbol{\lambda}}^{-1\prime} \hat{\boldsymbol{\lambda}}^{-1}}$ ... leading to $L_{\text{OLS}}(\hat{\mathbf{Z}})$.

A new class $\mathcal{L}$ of linear prediction scores

$\mathcal{L}$ contains PLS prediction scores

$$L_{\mathsf{PLS}}(\hat{\mathbf{Z}}, m) \quad \equiv \quad \boldsymbol{h}'_m \xi(\hat{\mathbf{Z}}).$$

with

- $\boldsymbol{h}_{m=1} = (1/\sqrt{p})\mathbf{1}_p$ ... leading to $L_{\mathsf{N}}(\hat{\mathbf{Z}})$
- $\boldsymbol{h}_{m=\min(n-1,p)} = \hat{\boldsymbol{\lambda}}^{-1}/\sqrt{\hat{\boldsymbol{\lambda}}^{-1\prime}\hat{\boldsymbol{\lambda}}^{-1}}$ ... leading to $L_{\mathsf{OLS}}(\hat{\mathbf{Z}})$.

Note : PCR and Lasso prediction scores do not belong to $\mathcal{L}$.

Optimal prediction within $\mathcal{L}$

The optimal vector $\boldsymbol{h}$ depends on $\boldsymbol{\lambda}$ and γ :

$$\boldsymbol{h}_{\text{opt}} = \left\{ \text{Var}(\xi(\hat{\boldsymbol{Z}})) \right\}^{-1} \text{Cov}(\xi(\hat{\boldsymbol{Z}}), Y).$$

where

$$\begin{aligned} \text{Var}(\xi(\hat{\boldsymbol{Z}})) &= \left(\boldsymbol{D}_{\boldsymbol{\lambda}} + \frac{\gamma\gamma'}{\sigma_y^2}\right) \odot \left(\gamma\gamma'\right) + o(n), \\ \text{Cov}\left\{\xi(\hat{\boldsymbol{Z}}), Y\right\} &= \gamma^{\odot 2} + o(n). \end{aligned}$$

Implemented in `R` package `AdaptiveRegression` available at `https://github.com/fhebert`.

Optimal prediction within $\mathcal{L}$

A toy simulation study $(n = 20,\ p = 19)$ [Witten and Tibshirani, 2009]

- $X_j,\ j = 1, \dots, 10$ are equicorrelated with $\rho = 0.9$
- $X_j,\ j = 11, \dots, 19$ are mutually independent and independent of $X_j,\ j = 1, \dots, 10$.
- Two scenarios for the association signal :
  - Sc. 1 : $\beta_j = j,\ j = 1, \dots, 10$ and $\beta_j = 0,\ j = 11, \dots, 19$
  - Sc. 2 : $\beta_j = 0,\ j = 1, \dots, 10$ and $\beta_j = 20 - j,\ j = 11, \dots, 19$

Optimal prediction within $\mathcal{L}$

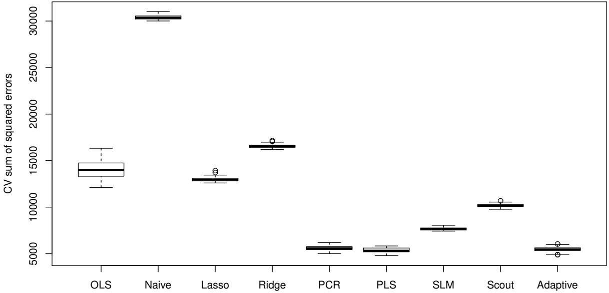
A toy simulation study $(n = 20,\ p = 19)$ [Witten and Tibshirani, 2009]

		Scenario 1	Scenario 2
Within $\mathcal{L}$	OLS	0.30 (0.17)	0.28 (0.17)
	Naive	0.79 (0.01)	0.23 (0.18)
	Ridge	0.73 (0.08)	0.55 (0.15)
	PLS	0.66 (0.30)	0.22 (0.27)
	Adaptive	0.76 (0.08)	0.52 (0.16)
Out of $\mathcal{L}$	Scout	0.76 (0.05)	0.54 (0.13)
	PCR	0.68 (0.27)	0.21 (0.25)

Optimal prediction within $\mathcal{L}$

Comparison study in a high-dimensional situation

- Li et al. (1996)'s data available in the R package `cggd` ;
- X : NIRS of samples of orange juice between 1100 and 2500 nm at 2 nm intervals ($n = 215, \ p = 700$) ;
- Y : is the concentration of saccharose.



Tentative conclusion

Three take-home messages

- To whiten or not to whiten is an ill-posed question
- The best handling depends on the true association signal and the dependence pattern
- Handling of dependence is not only a high-dimensional issue

Things I have not said :

- Estimating the optimal $\boldsymbol{h}$ raises numerical issues ;
- $\mathcal{L}$ is a general framework to derive exact optimization of hyperparameters for Ridge and PLS.