Application of the Lasso on genomic data
Part 3

Franck Picard*

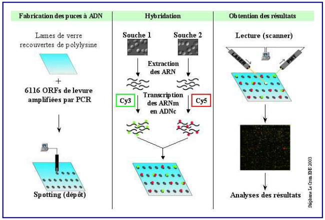
*Laboratoire de Biométrie et Biologie Évolutive, Univ. Lyon 1

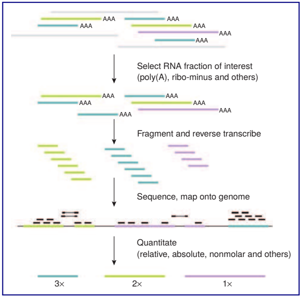
ECAS, High Dimensional Statistics Course, October 2017

Genomics and high dimensional statistics

- 1990-2000: DNA microarrays
- 2000-: massive parallel sequencing: RNA-Seq, chIP-Seq
- NGS : next generation sequencing

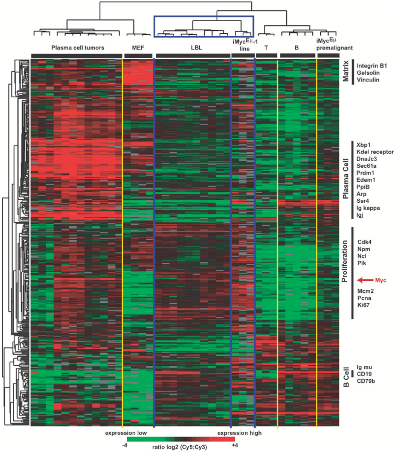
Genome wide molecular portraits of cells


Fabrication des puces à ADN
Hybridation
Obtention des résultats
Lames de verre recouvertes de polylysine
+
6116 ORFs de levure amplifiées par PCR
Spotting (dépôt)
Souche 1
Souche 2
Extraction des ARN
Cy3
Transcription des ARNm en ADNc
Cy5
Lecture (scanner)
LASER
PMT
Analyses des résultats
Stéphane Le Crom ENS 2003


AAA
Select RNA fraction of interest (poly(A), ribo-minus and others)
Fragment and reverse transcribe
Sequence, map onto genome
Quantitate (relative, absolute, nonmolar and others)
3×
2×
1×

The high throughtput point of view

- Before high throughput technologies, gene expression was quantified gene by gene
- High throughput technologies completely changed the perspectives of biologists
- In one experiment, one has access of the measurements of all transcripts in a cell (tissue)



Structure of datasets

	statut	exon_1	exon_2	$\dots$	exon_p	Age	Sexe	Glycemia
$i = 1$	0	10000	50		0	38	F	0.8
$i = 2$	1	10000	30		1	15	M	0.2
$\vdots$								
$i = N$	1	20000	25		3	90	F	1.5

For each individual

- Status (discrete/continuous)
- gene expression measurements (counts ou continuous)
- Clinical data

Goal

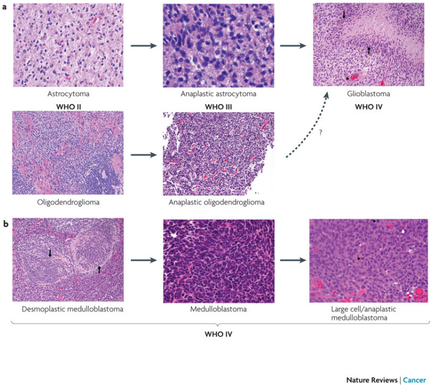
Explain the variations of a given response with gene expression measurements.

Example of statistical tasks

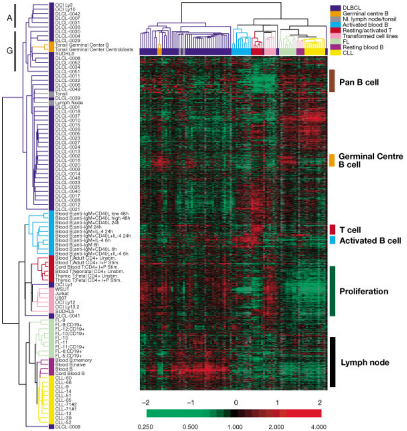
- Experimental Design
- Differential Analysis (multiple testing)
- Unsupervised Classification (clustering)
- Phenotype prediction (supervised)

Standard statistical tasks but need to be revisited to account for high dimension

New classifications and personalized medicine



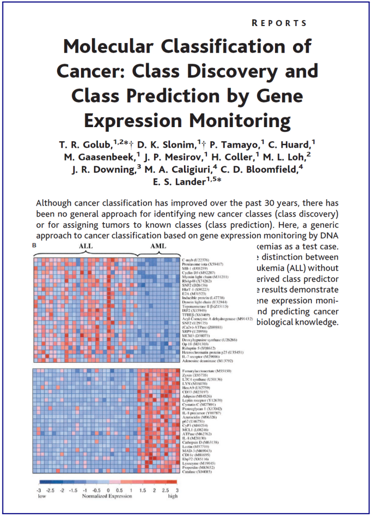
Classification based on histological data



Classification on molecular data

Towards new predictions ?

- In 1999 an article proposes to predict the molecular status of leukemia patients using genomic signatures
- The number of individuals is 38 for 6817 genes !
- Methodological developments for genomic signatures and prediction



REPORTS

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

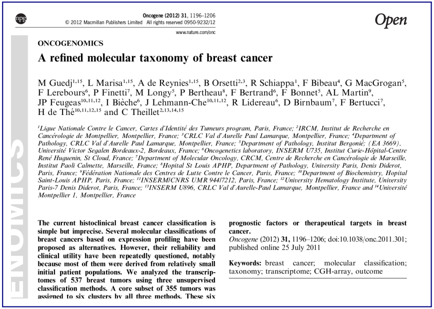
Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA

kemias as a test case.
e distinction between
ukemia (ALL) without
erived class predictor
e results demonstrate
ene expression moni-
nd predicting cancer
biological knowledge.

Data set: Breast Cancer Relapse

- level expression of 54613 genes for 294 patients affected by breast cancer.
- Y: relapse after 5 years (binary)
- 214 patients without relapse and 80 with a relapse.
- Preselection of 5000 genes
- What is the (ada)Lasso estimation of this data, estimation ? Prediction

▸ Lasso on genomic data



Oncogene (2012) 31, 1196–1206
© 2012 Macmillan Publishers Limited All rights reserved 0950-9232/12
www.nature.com/onc

Open

ONCOGENOMICS

A refined molecular taxonomy of breast cancer

M Guedj[1,15], L Marisa[1,15], A de Reynies[1,15], B Orsetti[2,3], R Schiappa[1], F Bibeau[4], G MacGrogan[5], F Lerebours[6], P Finetti[7], M Longy[5], P Bertheau[8], F Bertrand[6], F Bonnet[5], AL Martin[9], JP Feugeas[10,11,12], I Bièche[6], J Lehmann-Che[10,11,12], R Lidereau[6], D Birnbaum[7], F Bertucci[7], H de Thé[10,11,12,15] and C Theillet[2,13,14,15]

[1]Ligue Nationale Contre le Cancer, Cartes d'Identité des Tumeurs program, Paris, France; [2]IRCM, Institut de Recherche en Cancérologie de Montpellier, Montpellier, France; [3]CRLC Val d'Aurelle Paul Lamarque, Montpellier, France; [4]Department of Pathology, CRLC Val d'Aurelle Paul Lamarque, Montpellier, France; [5]Department of Pathology, Institut Bergonié; (EA 3669), Université Victor Segalen Bordeaux-2, Bordeaux, France; [6]Oncogenetics laboratory, INSERM U735, Institut Curie-Hôpital-Centre René Huguenin, St Cloud, France; [7]Department of Molecular Oncology, CRCM, Centre de Recherche en Cancérologie de Marseille, Institut Paoli Calmette, Marseille, France; [8]Hopital St Louis APHP, Department of Pathology, University Paris, Denis Diderot, Paris, France; [9]Fédération Nationale des Centres de Lutte Contre le Cancer, Paris, France; [10]Department of Biochemistry, Hopital Saint-Louis APHP, Paris, France; [11]INSERM/CNRS UMR 944/7212, Paris, France; [12]University Hematology Institute, University Paris-7 Denis Diderot, Paris, France; [13]INSERM U896, CRLC Val d'Aurelle-Paul Lamarque, Montpellier, France and [14]Université Montpellier 1, Montpellier, France

The current histoclinical breast cancer classification is simple but imprecise. Several molecular classifications of breast cancers based on expression profiling have been proposed as alternatives. However, their reliability and clinical utility have been repeatedly questioned, notably because most of them were derived from relatively small initial patient populations. We analyzed the transcriptomes of 537 breast tumors using three unsupervised classification methods. A core subset of 355 tumors was assigned to six clusters by all three methods. These six

prognostic factors or therapeutical targets in breast cancer.

Oncogene (2012) **31,** 1196–1206; doi:10.1038/onc.2011.301; published online 25 July 2011

Keywords: breast cancer; molecular classification; taxonomy; transcriptome; CGH-array, outcome

(Sparse) PCA to visualize the data

- $\mathbf{X}_{[n \times p]}$ measurements of p genes over n individuals (centered)
- Find a K dimensional subspace to represent the data
- Define $\mathbf{U}_{n \times K}$ the coordinates of the individuals in the new space
- Define $\mathbf{V}_{p \times K}$ the coordinates of the variables in the new space (loadings)

Approximate $\mathbf{X} \simeq \mathbf{UV}'$ by linear projection

Principal components

- $\mathbf{t}_k$ is a linear combination of the observed variables

$$\mathbf{t}_k = \mathbf{X}\mathbf{w}_k$$

- $\mathbf{w}_k \in \mathbb{R}^p$: contributions of the variables to the component
- The objective function is the empirical variance of the components $\widehat{\mathbb{V}}(\mathbf{t}_k)$ (under orthogonality constraints)

$$\widehat{\mathbf{w}}_k = \underset{\mathbf{w}_k \in \mathbb{R}^p}{\arg} \max \left\{ (\mathbf{w}_k \mathbf{X})' \mathbf{X} \mathbf{w}_k \right\}$$

- The solution is explicit and is given by the K eigen vectors of the empirical variance of $\mathbf{X}$ (centered)

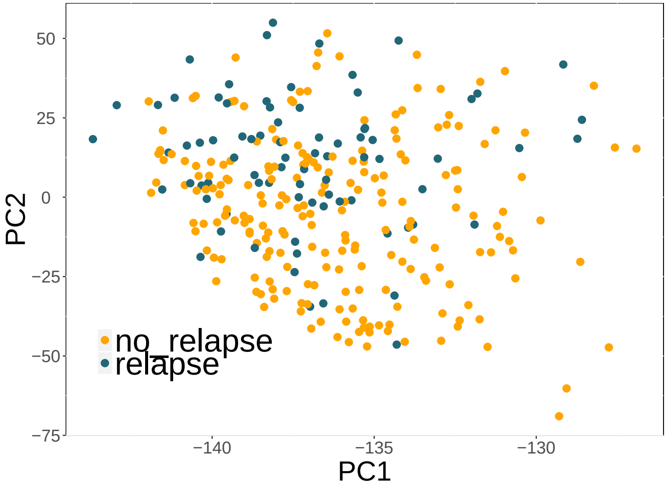
Sparse PCA ▸ Sparse PCA coding

- Some variables may contribute poorly to components (Ex of Sparse PCA paper)
- $\mathbf{w}_k$ are assumed to be sparse

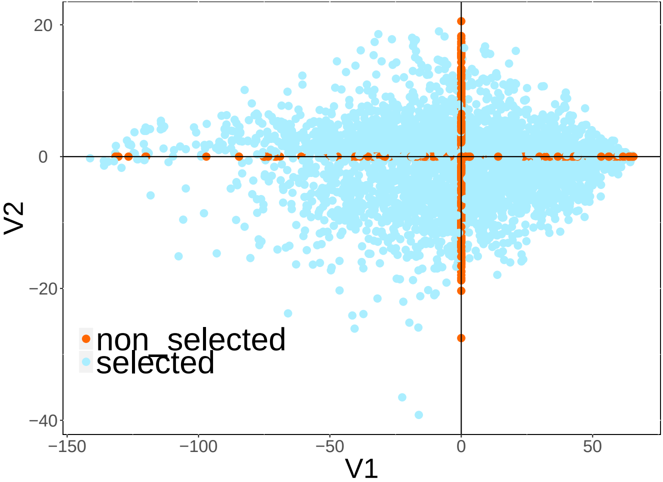
$$\widehat{\mathbf{w}}_k(\lambda) = \underset{\mathbf{w}_k \in \mathbb{R}^p}{\arg} \max \left\{ (\mathbf{w}_k \mathbf{X})' \mathbf{X} \mathbf{w}_k - \lambda \textstyle\sum_k \|\mathbf{w}_k\|_1 \right\}$$

- Also know as sparse coding or sparse matrix factorization
- Need to calibrate the hyperparameter by cross validation

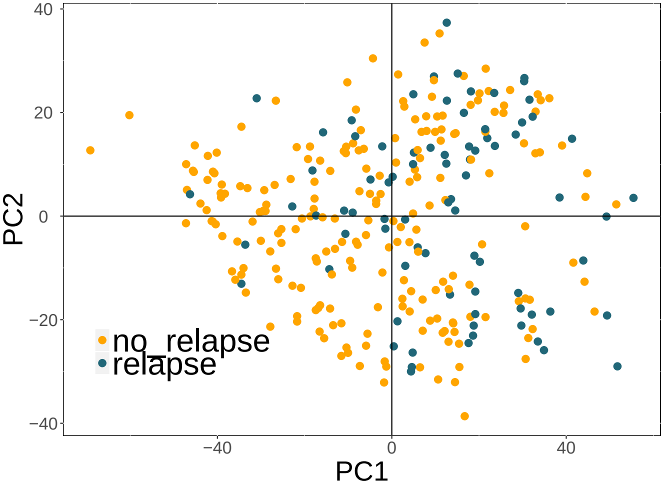
(Sparse) PCA to visualize the individuals



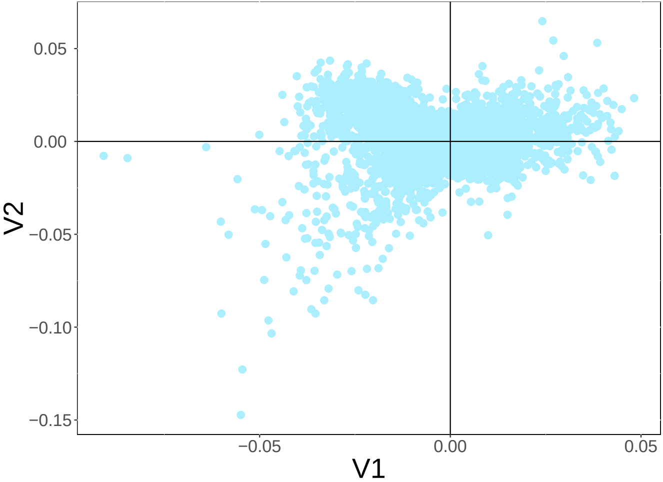
(Sparse) PCA to visualize the variables



(Sparse) PCA to visualize the individuals



(Sparse) PCA to visualize the variables



(Sparse) PLS for supervised dimension reduction

‣ Sparse PLS coding

- What if there is a response vector $\mathbf{Y}$?
- Instead of reduction dimension based on $\widehat{\mathbb{V}}(\mathbf{t}_k)$, use $\widehat{\mathbb{C}ov}(\mathbf{t}_k, \mathbf{Y})$

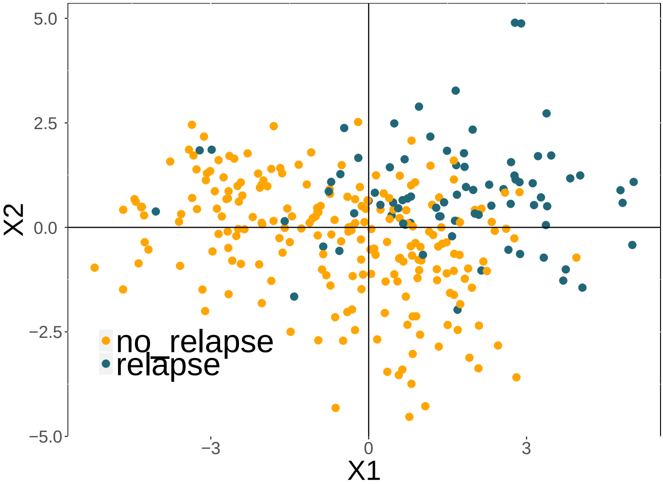
$$\widehat{\mathbf{w}}_k(\lambda) = \underset{\mathbf{w}_k \in \mathbb{R}^p \|\mathbf{w}_k\|_2^2 = 1}{\arg\max} \left\{ (\mathbf{w}_k \mathbf{X})' \mathbf{Y} - \lambda \sum_k \|\mathbf{w}_k\|_1 \right\}$$

- Can be reformulated as a regression problem: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$
- Considering $\mathbf{T} = \mathbf{XW}$ the matrix of principal components, PLS performs a regression of $\mathbf{Y}$ on $\mathbf{T}$

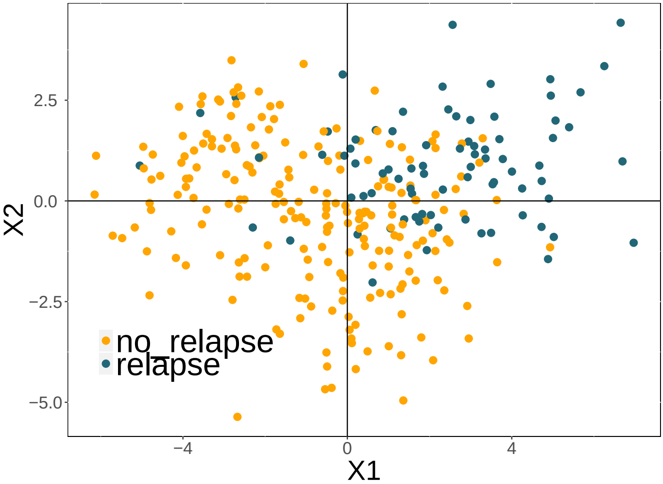
$$\mathbf{Y} = \mathbf{T}\gamma + \widetilde{\mathbf{E}}, \text{ with } \widehat{\beta} = \mathbf{W}\widehat{\gamma}$$

- Can be adapted to logistic regression ($\mathbf{Y}$ binary, tricky computations, *sparse logistic PLS paper*).

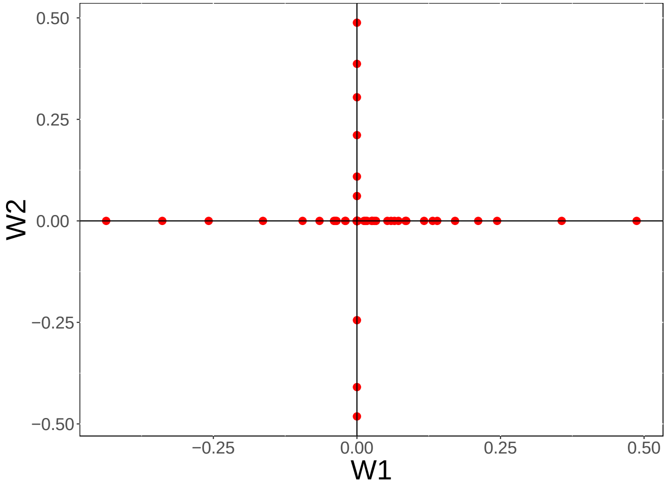
(Sparse) PLS to visualize the individuals



(Sparse) PLS to visualize the individuals



(Sparse) PLS to visualize the variables


0.50
0.25
0.00
−0.25
−0.50
W2
−0.25
0.00
0.25
0.50
W1