Machine learning for patient stratification from genomic data

Jean-Philippe Vert

Neurospin, November 20, 2017
Goal

Personalized Cancer Therapy

1. Molecular Profiling
2. Prognostic Markers
   - Markers predictive of drug sensitivity/resistance
   - Markers predictive of adverse events

https://pct.mdanderson.org
Mathematical model

- Patients with VS without relapse in 5 years
- $n (=19)$ patients $>> p (=2)$ markers
Mathematical model

- Patients with VS without relapse in 5 years
- $n (=19)$ patients $>> p (=2)$ markers
Mathematical model

- Patients with VS without relapse in 5 years
- \( n = 19 \) patients \( \gg p = 2 \) markers
Mathematical model

- Patients with VS without relapse in 5 years
- \( n (=19) \) patients \( \gg \) \( p (=2) \) markers
Real data: $n \ll p$

- Gene expression
- Somatic mutations

- $n = 10^2 \sim 10^4$ (patients)
- $p = 10^4 \sim 10^7$ (genes, mutations, copy number, ...)
- Data of various nature (continuous, discrete, structured, ...)
- Data of variable quality (technical/batch variations, noise, ...)

Data of various nature (continuous, discrete, structured, ...)
Data of variable quality (technical/batch variations, noise, ...)
Consequence: limited accuracy

Breast cancer prognosis competition, \( n = 2000 \) (Bilal et al., 2013)

- **C**: 16 standard clinical data (age, tumor size, ...)
- **M**: 80k molecular features (gene expression, DNA copy number)
Consequence: unstable biomarker selection

- 70 genes (Nature, 2002)
- 76 genes (Lancet, 2005)
- 3 genes in common

van ’t Veer et al. (2002); Wang et al. (2005)
Some research directions

- Regularize and incorporate prior knowledge
- Find a better representation
Outline

1. Regularize

2. Change representation
1. Regularize

2. Change representation
Typical problem

- $n$ samples (patients), $p$ features (genes)
- $X \in \mathbb{R}^{n \times p}$ gene expression profile of each patient
- $Y \in \mathcal{Y}^n$ survival information of each patient
- Fit a linear model for a sample $x \in \mathbb{R}^p$:

$$f(x) = \beta^\top x = \sum_{i=1}^{p} \beta_i x_i$$

- Standard methods (least squares or logistic regression) won’t work because $n < p$
Regularized linear models

In high dimension, estimate $\beta$ by solving

$$\min_{\beta \in \mathbb{R}^p} R(Y, X\beta) + \lambda J(\beta),$$

where

- $R(Y, X\beta)$ is an empirical risk to measures the fit to the training data
- $J(\beta)$ is a penalty to control the complexity of the model
- $\lambda > 0$ is a regularization parameter
Standard regularizations

\[
\min_{\beta \in \mathbb{R}^p} R(Y, X\beta) + \lambda J(\beta)
\]

where

- **Lasso**: \( J(\beta) = \|\beta\|_1 \) for gene selection.
- **Ridge**: \( J(\beta) = \|\beta\|_2^2 \) to address \( n \gg m \).
- **Elastic net**: \( J(\beta) = \alpha \|\beta\|_2^2 + (1 - \alpha) \|\beta\|_1 \)
Which regularization is the best?

- **Feature selection** (lasso, t-tests, ...) is popular, it leads to a limited set of genes that form a molecular signatures
- **Ridge** is less interpretable but often leads to better performance... e.g., breast cancer prognosis ($n = 286$):

![Influence of signature size on breast cancer prognosis performance.](image)
Adding prior knowledge: network-based regularizations

\[ G = (\mathcal{V}, \mathcal{E}) \] a graph of genes (PPI, metabolic, signaling, regulatory network...)

Prior knowledge:
- \( \beta \) should be "smooth" on the graph?
- Selected genes should be connected?
Examples of network-based regularizations

\[ J_G(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \]  
(Rapaport et al., 2007)

\[ J_G(\beta) = a \| \beta \|_1 + (1 - a) \sum_{i \sim j} (\beta_i - \beta_j)^2 \]  
(Li and Li, 2008)

\[ J_G(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j \alpha_i^2 + \alpha_j^2 \leq 1} \alpha^\top \beta \]  
(Jacob et al., 2009)

\[ J_G(\beta) = a \| \beta \|_1 + (1 - a) \sum_{i \sim j} |\beta_i - \beta_j| \]  
(Hoefling, 2010)
Gene selection with the graph lasso

\[
\Omega(\beta) = \sup_{\alpha \in \mathbb{R}^p: \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta
\]

Jacob et al. (2009)
BC prognosis: Lasso signature (accuracy 0.61)

Jacob et al. (2009)
BC prognosis: Graph Lasso signature (accuracy 0.64)

Jacob et al. (2009)
Smoothness regularization and Fourier transform

- "Connected genes have similar weights" (Rapaport et al., 2007; Li and Li, 2008)
  \[
  J_G(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2
  \]

- No feature selection

- Reinterpretation in the Fourier domain (Rapaport et al., 2007):
  \[
  \sum_{i \sim j} (\beta_i - \beta_j)^2 = \sum_{i=1}^{p} \lambda_i \hat{\beta}_i^2
  \]
  where
  - \( \hat{\beta}_i \) is the \( i \)-th Fourier coefficient of \( \beta \)
  - \( \lambda_i \) is the \( i \)-th frequency
  - "\( \beta \) has little energy at high frequency" and is therefore smooth on the graph
"Connected genes have similar weights" (Rapaport et al., 2007; Li and Li, 2008)

\[ J_G(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \]

- No feature selection
- Reinterpretation in the Fourier domain (Rapaport et al., 2007):

\[ \sum_{i \sim j} (\beta_i - \beta_j)^2 = \sum_{i=1}^{p} \lambda_i \hat{\beta}_i^2 \]

where
- \( \hat{\beta}_i \) is the \( i \)-th Fourier coefficient of \( \beta \)
- \( \lambda_i \) is the \( i \)-th frequency

"\( \beta \) has little energy at high frequency" and is therefore smooth on the graph
Graph Fourier transform $\hat{\beta}$?

- Eigenvectors $U$ of the graph Laplacian matrix form the Fourier basis:
  $$\hat{\beta} = U^T \beta$$

- Eigenvalues $\Lambda = (0 = \lambda_1 \leq \ldots \leq \lambda_p)$ represent the "frequencies" of the Fourier basis.

![Graph Laplacian matrix representation](image-url)
Graph Fourier transform $\hat{\beta}$?

- Eigenvectors $U$ of the graph Laplacian matrix form the Fourier basis:
  \[
  \hat{\beta} = U^\top \beta
  \]

- Eigenvalues $\Lambda = (0 = \lambda_1 \leq \ldots \leq \lambda_p)$ represent the "frequencies" of the Fourier basis
Graph Fourier transform $\hat{\beta}$?

- Eigenvectors $U$ of the graph Laplacian matrix form the Fourier basis:
  \[ \hat{\beta} = U^T \beta \]

- Eigenvalues $\Lambda = (0 = \lambda_1 \leq \ldots \leq \lambda_p)$ represent the "frequencies" of the Fourier basis.

![Diagram](image.png)
Eigenvectors $U$ of the graph Laplacian matrix form the Fourier basis:

$$\hat{\beta} = U^T \beta$$

Eigenvalues $\Lambda = (0 = \lambda_1 \leq \ldots \leq \lambda_p)$ represent the "frequencies" of the Fourier basis.
Graph Fourier transform $\hat{\beta}$?

- Eigenvectors $U$ of the graph Laplacian matrix form the Fourier basis:
  $$\hat{\beta} = U^\top \beta$$

- Eigenvalues $\Lambda = (0 = \lambda_1 \leq \ldots \leq \lambda_p)$ represent the "frequencies" of the Fourier basis.
Graph Fourier transform $\hat{\beta}$?

- Eigenvectors $U$ of the graph Laplacian matrix form the Fourier basis:
  $$\hat{\beta} = U^T \beta$$

- Eigenvalues $\Lambda = (0 = \lambda_1 \leq \ldots \leq \lambda_p)$ represent the "frequencies" of the Fourier basis.

![Graph](image)
Graph Fourier transform $\hat{\beta}$?

- Eigenvectors $U$ of the graph Laplacian matrix form the Fourier basis:
  \[
  \hat{\beta} = U^T \beta
  \]
- Eigenvalues $\Lambda = (0 = \lambda_1 \leq \ldots \leq \lambda_p)$ represent the "frequencies" of the Fourier basis

![Graph Laplacian matrix](image.png)
Graph Fourier transform $\hat{\beta}$?

- Eigenvectors $U$ of the graph Laplacian matrix form the Fourier basis:
  $$\hat{\beta} = U^T \beta$$

- Eigenvalues $\Lambda = (0 = \lambda_1 \leq \ldots \leq \lambda_p)$ represent the "frequencies" of the Fourier basis

Lambda = 4.2

lambda = 3.5

\begin{itemize}
  \item [2] 4 6 8
  \item [-0.4 -0.2 0.0 0.2 0.4]
\end{itemize}
Graph Fourier transform $\hat{\beta}$?

- Eigenvectors $U$ of the graph Laplacian matrix form the Fourier basis:
  $$\hat{\beta} = U^T \beta$$

- Eigenvalues $\Lambda = (0 = \lambda_1 \leq \ldots \leq \lambda_p)$ represent the "frequencies" of the Fourier basis.
Rapaport et al. (2007) extends
\[
\sum_{i \sim j} (\beta_i - \beta_j)^2 = \sum_{i=1}^{p} \lambda_i \beta_i^2
\]
to
\[
\sum_{i=1}^{p} \phi(\lambda_i) \beta_i^2
\]
for \( \phi : \mathbb{R}^+ \to \mathbb{R}^+ \) non-decreasing.

Example: \( \phi(\lambda) = \exp(-\gamma \lambda) \) linked to the diffusion kernel on the graph.
Fourier vs wavelets

Fourier

Localized in frequency

Wavelets

Localized in frequency AND space
Fig. 5. Spectral graph wavelets on cerebral cortex, with $K = 50$, $J = 4$ scales. (a) ROI with white wavelets centered, (b) scaling function, (c) – (f) wavelets, scales 1–4.

Fig. 6. Spectral graph wavelets on lake Geneva domain (spatial map (a), contour plot (c)); compared with truncated wavelets from graph corresponding to complete mesh (spatial map (b), contour plot (d)). Note that the graph wavelets adapt to the geometry of the domain.

(Hammond et al., 2011)
Stability performance of gene selection related to breast cancer survival, estimated over 100 random experiments. The black dotted curve denotes random selection. From Jiao and Vert (to appear)
Connectivity performance of gene selection related to breast cancer survival, where special marks correspond to the number tuned by cross-validation. The black dotted curve denotes random selection. From Jiao and Vert (to appear)
Regularization is needed in high dimension
While gene selection is popular, alternatives exist which often work better
Different strategies to include prior knowledge
- structured feature selection (variants of lasso)
- smoothness (in the Fourier domain)
- wavelet decomposition (frequency/localization)
1. Regularize

2. Change representation
Between-sample variability: batch effect, drift over time, ...

Typical pre-processing: Quantile normalization per sample

Only the relative ordering of features within each sample is used
Learning with permutations

- Represent each sample $x \in \mathbb{R}^p$ by the ranks of genes $\sigma \in S_p$
- The symmetric group $S_p$ is the set of permutations of $\{1, \ldots, p\}$
Represent $x \in \mathbb{R}^p$ by $\Pi_x \in \mathbb{R}^{p \times p}$ with

$$[\Pi_x]_{ij} = \begin{cases} 1 & \text{if } x_j \text{ has rank } i, \\ 0 & \text{otherwise}. \end{cases}$$
Learn a linear model

\[ f(x) = \langle M, \Pi_x \rangle_{\text{Frobenius}} = \text{trace}(M^\top \Pi(x)) \]

Constrain \( \text{rank}(M) = 1 \)

This is equivalent to quantile normalization, where the target quantile function is jointly optimized: we call is supervised quantile normalization, a.k.a. **SUQUAN** (Le Morvan and Vert, 2017)
Proof: from $\Pi_x$ to SUQUAN

- QN with target quantile $f \in \mathbb{R}^p$ is $\Pi_x f$.
- Learning linear model $f(u) = w^\top u + b$ on QN-transformed data while optimizing $f$ is:

$$
\min_{w,b,f} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_i \left( w^\top \Phi_f(x_i) + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}
$$

$$
= \min_{w,b,f} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell \left( w^\top \Pi_x f + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}
$$

$$
= \min_{w,b,f} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell \left( \langle wf^\top, \Pi x_i \rangle_{Fro} + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}
$$

- A particular **linear model** to estimate a rank-1 matrix $M = wf^\top$
- Non-convex
- Local optimum found by alternatively optimizing $f$ and $w$
Results: gene expression data

<table>
<thead>
<tr>
<th></th>
<th>RAW</th>
<th>RMA</th>
<th>CAUCHY</th>
<th>EXP.</th>
<th>UNIF.</th>
<th>GAUS.</th>
<th>MEDIAN</th>
<th>SVD</th>
<th>BND</th>
<th>SPAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE1456</td>
<td>65.94</td>
<td>68.73</td>
<td>59.56</td>
<td>68.86</td>
<td>68.72</td>
<td>69.00</td>
<td>69.06</td>
<td>57.60</td>
<td>71.44</td>
<td>69.60</td>
</tr>
<tr>
<td>GSE2034</td>
<td>74.52</td>
<td>75.42</td>
<td>61.91</td>
<td>74.53</td>
<td>75.22</td>
<td>76.45</td>
<td>74.92</td>
<td>52.61</td>
<td>70.50</td>
<td>76.11</td>
</tr>
<tr>
<td>GSE2990</td>
<td>57.01</td>
<td>60.43</td>
<td>54.72</td>
<td>61.25</td>
<td>56.25</td>
<td>58.66</td>
<td>59.72</td>
<td>52.51</td>
<td>59.22</td>
<td>59.94</td>
</tr>
<tr>
<td>GSE4922</td>
<td>58.52</td>
<td>58.86</td>
<td>55.24</td>
<td>58.81</td>
<td>55.66</td>
<td>60.01</td>
<td>59.18</td>
<td>52.39</td>
<td>61.82</td>
<td>61.41</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>64.00</td>
<td>65.86</td>
<td>57.86</td>
<td>65.86</td>
<td>63.96</td>
<td>66.03</td>
<td>65.72</td>
<td>53.78</td>
<td>65.75</td>
<td>66.77</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE2034</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSE4922</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Estimated quantile function: iteration=0
Estimated quantile function: iteration=1
Estimated quantile function: iteration=2
Another representation of permutations

One sample $x$  
$p$ features

Mapping $f(x)$  
$p(p-1)/2$ bits

$$
\Phi_{i,j}(x) = \begin{cases} 
1 & \text{if } x_i \leq x_j, \\
0 & \text{otherwise.}
\end{cases}
$$
Link with Kendall’s $\tau$ (Jiao and Vert, 2017)

Useful in practice (kernel methods)
Average performance on 10 microarray classification problems (Jiao and Vert, 2017).
Somatic mutations in cancer

Stratton et al. (2009)
Large-scale efforts to collect somatic mutations

- 3,378 samples with survival information from 8 cancer types
- downloaded from the TCGA / cBioPortal portals.

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Patients</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUAD (Lung adenocarcinoma)</td>
<td>430</td>
<td>20 596</td>
</tr>
<tr>
<td>SKCM (Skin cutaneous melanoma)</td>
<td>307</td>
<td>17 463</td>
</tr>
<tr>
<td>GBM (Glioblastoma multiforme)</td>
<td>265</td>
<td>14 750</td>
</tr>
<tr>
<td>BRCA (Breast invasive carcinoma)</td>
<td>945</td>
<td>16 806</td>
</tr>
<tr>
<td>KIRC (Kidney renal clear cell carcinoma)</td>
<td>411</td>
<td>10 609</td>
</tr>
<tr>
<td>HNSC (Head and Neck squamous cell carcinoma)</td>
<td>388</td>
<td>17 022</td>
</tr>
<tr>
<td>LUSC (Lung squamous cell carcinoma)</td>
<td>169</td>
<td>13 590</td>
</tr>
<tr>
<td>OV (Ovarian serous cystadenocarcinoma)</td>
<td>363</td>
<td>10 195</td>
</tr>
</tbody>
</table>
Replace $x \in \{0, 1\}^p$ by $\Phi(x) \in \{0, 1\}^p$, using a gene network as prior knowledge.

Enforce quantile normalization, i.e., after Netnorm, all patients $\Phi(x)$ have the same number of (pseudo-)mutations.
1. Add mutations for patients with few (less than \(k\)) mutations

2. Remove mutations for patients with many (more than \(k\)) mutations

In practice, \(k\) is a free parameter optimized on the training set, typically a few 100's.
Performance on survival prediction

Use Pathway Commons as gene network.

NSQN = Network Smoothing / Quantile Normalization (Hofree et al., 2013)
A good representation is worth a thousand ML algorithms

**Permutations** offer an interesting setting
- robust to various sources of noise
- amenable to machine learning (SUQUAN, Kendall kernel)

**Learning** representations is a hot topic (deep learning...
PROMOTION DU PROGRAMME

• Rencontre avec les directeurs, directeurs scientifiques et directeurs des études de l'ENS, Mines, ENSCP, ESPCI pour organiser la promotion du programme auprès des étudiants.

• Ecole des Mines : présentation d'ITI devant le conseil de l'enseignement ; 8 représentants d'élèves ; relance.

• ENSCP : mailing de présentation d'ITI aux 3A via la direction des études.

• ENS : mailing ; présentation directe auprès des étudiants (2 élèves présents) ; diffusion des plaquettes et du syllabus.

• ESPCI : mailing aux 3A + présentation d'ITI / rencontre avec les étudiants en présence de la directrice de la scolarité et directrice des relations entreprises ; 8 élèves présents.

• Contacts en cours avec les Ponts et l'ENS Lyon (en attente de réponse).

• Rencontre à l'ANRT avec le délégué général et la chef du service CIFRE : accord de communication sur le programme ITI via le site internet et de l'ANRT (rubrique "zoom sur.

• Promotion aux Rencontres Universités Entreprises (RUE).

• Echange avec Stéphane Mallat et réflexion sur la pertinence du programme tronc commun.

• Rencontre avec le responsable des relations internationales de la National Taiwan University à ESPCI : présentation d'ITI.

• Contact en cours pour visite d'entreprise.

• Articles dans les Echos.

• Création du logo PSL/ITI et du Diplôme Supérieur de Recherche et d'Innovation de PSL/ITI (DSRI), dépôt INPI en cours.

2/ MISE EN ŒUVRE OPERATIONNELLE DU PROGRAMME


