Machine learning for patient stratification from genomic data

Jean-Philippe Vert

IHES, March 9, 2018
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 \( >> \) \( p (=2) \) 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, ...)
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

Gene expression profiling predicts clinical outcome of breast cancer


* Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands
‡ Rosetta Inpharmatics, 17040 115th Avenue NE, Kirkland, Washington 98034.


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
Outline

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 $n < p$
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 ($n = 286$):

![Graph showing influence of signature size on breast cancer prognosis performance.](image)

**FIGURE 6.4** Influence of signature size on breast cancer prognosis performance. A regularised LR classifier using a signature of varying size is trained on the Wang expression dataset to predict relapse within 5 years. The genes in the signatures are selected either randomly, or by decreasing significance according to a t-test. The performance is estimated by 5-fold cross-validation, averaged over 10 repeats. In this example, it is better to keep all genes to train the classifier.

The notion of **genomic grade** to quantify tumour differentiation (Sotiriou et al., 2003; Loi et al., 2007). In addition to tumour differentiation assessment, this genomic grade was shown to be prognostic. Several prognostic molecular predictors have also been proposed, including the 76-gene MammaPrint signature developed at the Netherlands Cancer Institute in Amsterdam (van't Veer et al., 2002) and the 76-gene Rotterdam signature of Wang et al. (2005). Investigators from the University of Texas M. D. Anderson Cancer Center developed DLD30, a 30-gene signature to predict the response of a tumour to preoperative chemotherapies (Hess et al., 2006). The Oncotype DX assay combines the expression of 21 genes to evaluate the risk of relapse and the benefits of chemotherapy for patients with early-stage, lymph node-negative, ER+/HER2- breast cancers (Paik et al., 2006; Paik, 2007). Several of these molecular predictors have reached the level of clinical trials, and are now being tested on large cohorts of patients. We can already foresee their routine use in the clinics within few years.

6.3.6 **Pitfalls and challenges in biomarker discovery**

Although an attractive strategy to improve the performance of predictive modelling in high-dimension and simultaneously identify biologically relevant markers, the automatic data-driven identification of new markers remains challenging for several reasons.
Adding prior knowledge: network-based regularizations

\[ \mathcal{G} = (\mathcal{V}, \mathcal{E}) \] 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

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

Jacob et al. (2009)
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 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 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 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.
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:
  \[ \hat{\beta} = U^T \beta \]

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

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

Lambda = 3.6
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 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

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

to

\[
\sum_{i=1}^{p} \phi(\lambda_i) \hat{\beta}_i^2
\]

for \( \phi : \mathbb{R}^+ \to \mathbb{R}^+ \) non-decreasing.

Example: \( \phi(\lambda) = \exp(-\gamma \lambda) \) linked to the diffusion kernel on the graph.
1. Regularize

2. Change representation
Back to the data
From raw data to $X$

- Between-sample variability: batch effect, drift over time, ...
- Typical pre-processing: Quantile normalization per sample
Fix a target quantile $f \in \mathbb{R}^n$

Transform $x \in \mathbb{R}^p$ to $\Phi_f(x)$ such that:
- The ranking of entries in $x$ and $\Phi_f(x)$ are the same
- The distribution of entries in $\Phi_f(x)$ follows $f$

See also: images (Gonzalez and Woods, 2008), MRI scans (Shinohara et al., 2014), speech (Hilger and Ney, 2006)
How to choose a "good" target distribution?

- Gaussian distribution (mean=0, sd=1)
- Uniform distribution
- Bigaussian distribution

Quantile function transformations:
- Gaussian to Gaussian
- Uniform to Uniform
- Bigaussian to Bigaussian
From QN to supervised QN (Le Morvan and Vert, 2017)

Standard approaches: learn model after QN preprocessing:

1. **Fix** $f$ arbitrarily (typically, mean empirical quantile function)
2. QN all samples to get $\Phi_f(x_1), \ldots, \Phi_f(x_n)$
3. Learn a model on normalized data, e.g.:

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

**SUQUAN**: jointly learn $f$ and the model:

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

For $x \in \mathbb{R}^p$ let

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

Then

$$\Phi_f(x) = \Pi_x f$$
Linear SUQAN as rank-1 matrix regression

- Linear SUQUAN therefore solves

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

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

\[
= \min_{w,b,f} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell \left( < w f^\top, \Pi_{x_i} > \text{Frobenius} + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}
\]

- A particular linear model to estimate a rank-1 matrix \( M = w f^\top \)
- Each sample \( x \in \mathbb{R}^p \) is represented by the matrix \( \Pi_x \in \mathbb{R}^{p\times p} \)
- Non-convex
- Alternative optimization of \( f \) and \( w \) is easy
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></td>
<td>57.60</td>
<td>71.44</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></td>
<td>52.61</td>
<td>70.50</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></td>
<td>52.51</td>
<td>59.22</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></td>
<td>52.39</td>
<td>61.82</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></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></td>
<td>53.78</td>
<td>65.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>SVD</th>
<th>BND</th>
<th>SPAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE2034</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>5000</td>
<td>15000</td>
</tr>
<tr>
<td>GSE4922</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>5000</td>
<td>15000</td>
</tr>
</tbody>
</table>

GSE2034

- **median**
- **SUQUAN BND**
- **SUQUAN SPAV**

GSE4922

- **median**
- **SUQUAN BND**
- **SUQUAN SPAV**
Estimated quantile function: iteration=0
Estimated quantile function: iteration=1
Estimated quantile function: iteration=2
Remark: embedding $\mathbb{R}^n$ to $S_n$

- Remark: each sample $x \in \mathbb{R}^p$ was represented by the permutation of genes $\sigma \in S_p$
- Many other possibilities when we decide to embed data to the symmetric group $S_n$
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>
Patient stratification (unsupervised) from raw mutation profiles

✓ Non-Negative matrix factorisation (NMF)

✓ Desired behaviour:

\[
\begin{array}{cccc}
N = 2 & N = 3 & N = 4 & N = 5 & N = 6 \\
\end{array}
\]

✓ Observed behaviour:

\[
\begin{array}{cccc}
N = 2 & N = 3 & N = 4 & N = 5 & N = 6 \\
\end{array}
\]

Patients share very few mutated genes!
Survival prediction from raw mutation profiles

- Each patient is a **binary vector**: each gene is mutated (1) or not (2)
- Silent mutations are removed
- Survival model estimated with sparse survival SVM
- Results on 5-fold cross-validation repeated 4 times
Can we replace

\[ x \in \{0, 1\}^p \quad \text{with } p \text{ very large, very sparse} \]

by a representation with more information shared between samples

\[ \Phi(x) \in \mathcal{H} \]

that would allow better supervised and unsupervised classification?
Take

\[ \mathcal{H} = \left\{ x \in \{0, 1\}^p : \sum_{i=1}^{p} x_i = K \right\} \]

and use a gene network to transform \( x \) to \( \phi(x) \in \mathcal{H} \) by adding/removing mutations.
Add mutations for patients with few (less than $K$) mutations

Remove mutations for patients for many (more than $K$) mutations

In practice, $K$ is a free parameter optimized on the training set, typically a few 100's.
Network-based stratification of tumor mutations

Matan Hofree¹, John P Shen², Hannah Carter², Andrew Gross³ & Trey Ideker¹–³

¹Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. ²Department of Medicine, University of California, San Diego, La Jolla, California, USA. ³Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. Correspondence should be addressed to T.I. (tideker@ucsd.edu).

RECEIVED 14 FEBRUARY; ACCEPTED 12 AUGUST; PUBLISHED ONLINE 15 SEPTEMBER 2013; DOI:10.1038/NMETH.2651

Network smoothing:

- Gene
- Gene-gene interaction
- Patient genotype 1
- Patient genotype 2
- Co-occurrence of genotype 1 and 2

Network-based stratification of tumor mutations:

- Patients
- Genes
- Network influence constraint
- Cluster prototypes
- Matrix smoothing
- Network regions

For further details, see Online Methods. To evaluate the impact of network smoothing, the mutational activity of a gene is a continuous value reflected in the intensity of yellow or blue; genes with high scores in both patients appear in yellow (patient 1) and blue (patient 2) in the context of a gene interaction network. Following the matrix decomposition, network regions characteristic of each subtype and co-occurrence of genotype 1 and 2 are identified. Finally, to promote robust cluster assignments, consensus stratification of tumor mutations is derived from the consensus (majority) assignments of each tumor after 1,000 applications of the network-based stratification method.
Results: unsupervised classification

[Graphs showing log-rank statistic for various types of cancer (LUSC, HNSC, OV, BRCA, KIRC, GBM, SKCM, LUAD) with different numbers of subtypes (2-6) and survival probability over time (months)]

- LUSC
- HNSC
- OV
- BRCA
- KIRC
- GBM
- SKCM
- LUAD

Number of subtypes

Survival probability

Time (months)
Use Pathway Commons as gene network.
NSQN = Network Smoothing / Quantile Normalization (Hofree et al., 2013)
QN matters...

Both NetNorm and NSQN transforms follow a 2-step approach:

1. Smooth the raw data onto the gene network (NS)
2. Quantile normalize the smoothed profile (QN)
Conclusion

- Learning from genomic data is **challenging**
- **Regularization** is needed in high dimension
- A good **representation** is worth a thousand learning algorithms
- Subtle **interplay** between biology and math/CS
- **Impact** on the final quality/performance of the model
- Recent trend: **learn** the representation
1/ 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 intern 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 l'ESPCI: présentation d'ITI

- Contact en cours pour visite d'entreprise

- Article dans les Échos

- 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


NetNorM and NSQN benefit from biological information in the gene network.

Comparison with 10 randomly permuted networks:
Selected genes represent "true" or "proxy" mutations

Genes selected in at least 50% of the cross-validated sparse SVM model
Proxy mutations encode both total number of mutations and local mutational burden

Figure 4 – Analysis of predictive genes. (a) Comparison of survival prediction performances according to patients' mutation rate for LUAD. Three different representations of the mutations are used to perform survival prediction using a ranking SVM: raw (the raw binary mutation data), NSQN (network smoothing with quantile normalisation) and NetNorM. NSQN and NetNorM are applied with Pathway Commons as gene-gene interaction network. Performances for half of the patients with fewer (resp. more) mutations are derived from the predictions made using the whole dataset. (b) Scatter plot of the correlation between the total number of mutations across patients and the number of mutated neighbours of a gene across patients (x-axis) against the degree of a gene (y-axis). This plot was generated using the raw mutation data for LUAD and Pathway Commons. (c) Scatter plot of the total number of mutations in a patient (x-axis) against the number of mutated neighbours of KHDRBS1 in a patient (y-axis). Only patients with less that $k_{\text{med}} = 295$ mutations are shown, where $k_{\text{med}}$ is the median value of $k$ learned across cross-validation folds. Red (resp. blue) indicate patients mutated (resp. non mutated) in KHDRBS1 after processing with NetNorM using $k = k_{\text{med}}$. The black line was fit by linear regression and by definition indicates the expected number of mutated neighbours of KHDRBS1 given the mutation rate of a patient. The plot was generated using the LUAD dataset with Pathway Commons.

KHDRBS1: a member of the K homology domain-containing, RNA-binding, signal transduction-associated protein family
Adding good old clinical factors

Combination by averaging predictions
QN after network smoothing
\[ \Phi_{i,j}(x) = \begin{cases} 
1 & \text{if } x_i \leq x_j, \\
0 & \text{otherwise.} 
\end{cases} \]
Geometry of the embedding

For any two permutations $\sigma, \sigma' \in S_n$:

- **Inner product**
  $$\Phi(\sigma)\top \Phi(\sigma') = \sum_{1 \leq i \neq j \leq n} 1_{\sigma(i) < \sigma(j)} 1_{\sigma'(i) < \sigma'(j)} = n_c(\sigma, \sigma')$$

  $n_c = \text{number of concordant pairs}$

- **Distance**
  $$\| \Phi(\sigma) - \Phi(\sigma') \|^2 = \sum_{1 \leq i, j \leq n} (1_{\sigma(i) < \sigma(j)} - 1_{\sigma'(i) < \sigma'(j)})^2 = 2n_d(\sigma, \sigma')$$

  $n_d = \text{number of discordant pairs}$
Kendall and Mallows kernels (Jiao and Vert, 2017)

- The Kendall kernel is

\[ K_{\tau}(\sigma, \sigma') = n_c(\sigma, \sigma') \]

- The Mallows kernel is

\[ \forall \lambda \geq 0 \quad K_{\lambda}^{M}(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')} \]

**Theorem (Jiao and Vert, 2015, 2017)**

The Kendall and Mallows kernels are positive definite.

**Theorem (Knight, 1966)**

These two kernels for permutations can be evaluated in \( O(n \log n) \) time.

*Kernel trick useful with few samples in large dimensions*
Kondor and Barbarosa (2010) proposed the diffusion kernel on the Cayley graph of the symmetric group generated by adjacent transpositions. Computationally intensive ($O(n^2n)$)

Mallows kernel is written as

$$K^\lambda_M(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')} ,$$

where $n_d(\sigma, \sigma')$ is the shortest path distance on the Cayley graph.

It can be computed in $O(n \log n)$
Average performance on 10 microarray classification problems (Jiao and Vert, 2017).