Learning from omics data

Jean-Philippe Vert

institutCurie
Together, let's beat cancer.

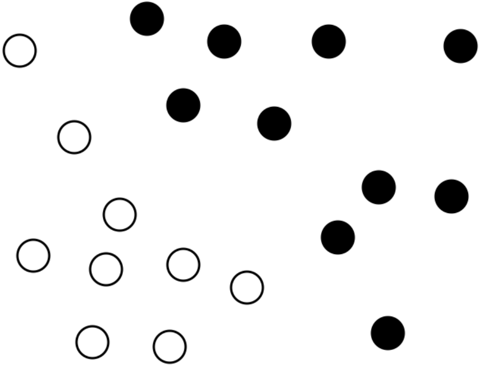
Computational Cancer Biology workshop, Simons Institute, Berkeley, Feb 5, 2016

Motivation

Also: diagnosis, prognosis, cell classification, drug response prediction, ...

Machine learning formulation

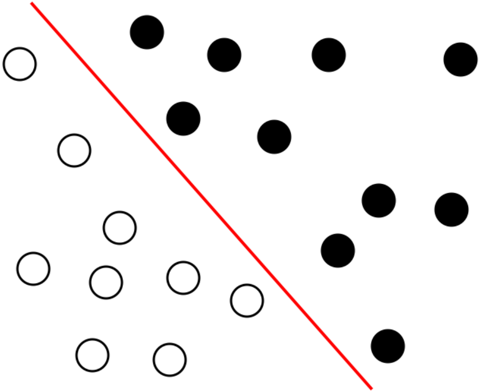
$n (= 19) >> p (= 2)$: easy



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

Machine learning formulation

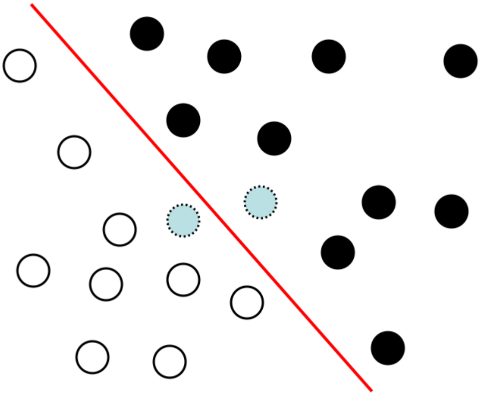
$n(= 19) >> p(= 2)$: easy



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

Machine learning formulation

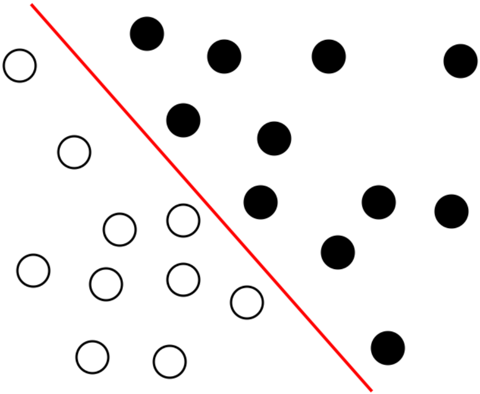
$n (= 19) >> p (= 2)$: easy



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

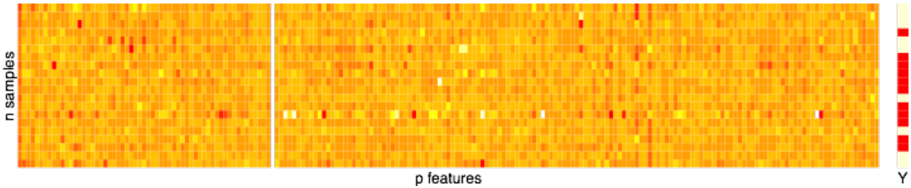
Machine learning formulation

$n (= 19) >> p (= 2)$: easy



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

*-omics challenge: $n << p$



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

Consequences: Accuracy drops, biomarker selection unstable

Can we replace the high-dimensional profile of a sample by a "simpler" representation, more amenable to statistical learning?

Outline

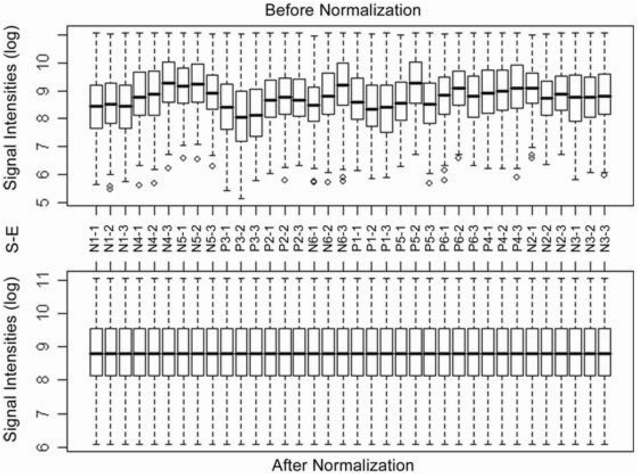
1. SUQUAN: Supervised full quantile normalization (w. Marine Le Morvan)
2. Learning from pairwise comparisons with the Kendall and Mallows kernels (w. Yunlong Jiao)

Outline

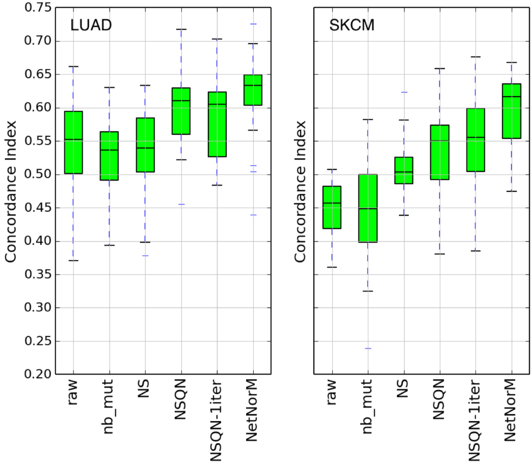
1. SUQUAN: Supervised full quantile normalization (w. Marine Le Morvan)

2. Learning from pairwise comparisons with the Kendall and Mallows kernels (w. Yunlong Jiao)

Full quantile normalization



Quantile normalization matters



(Marine's talk)

How to choose the target distributions?
Gaussian? Uniform? CDF of the data?

Learning the target distribution

- Let $f \in \mathbb{R}^p$ a non-decreasing target distribution (CDF)
- For $x \in \mathbb{R}^p$, let $\Phi_f(x) \in \mathbb{R}^p$ be the data after full quantile normalization with target distribution f
- Learn a (generalized) linear model over normalized data:

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

- SUQUAN: jointly learn f and (w, b):

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

SUQAN: supervised quantile normalization

- For $x \in \mathbb{R}^p$, let $\Pi_x \in \mathbb{R}^{p \times p}$ the permutation matrix of x's entries
- Quantile normalized x with target distribution f is:

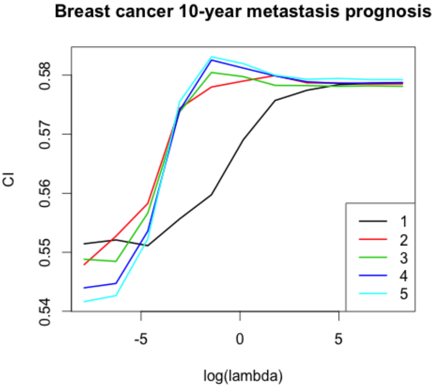
$$\Phi_f(x) = \Pi_x f$$

- SUQAN solves

$$\begin{aligned} & \min_{w,b,f} \frac{1}{n} \sum_{i=1}^{n} \ell\left(w^\top \Pi_{x_i} f + b\right) + \lambda \Omega(w) \\ = & \min_{w,b,f} \frac{1}{n} \sum_{i=1}^{n} \ell\left(< wf^\top, \Pi_{x_i} > + b\right) + \lambda \Omega(w) \end{aligned} \tag{1}$$

- A particular rank-1 matrix optimization, x is represented by Π_x
- Efficiently solved by alternatively optimizing f (isotonic GLM) and w

Results (preliminary)



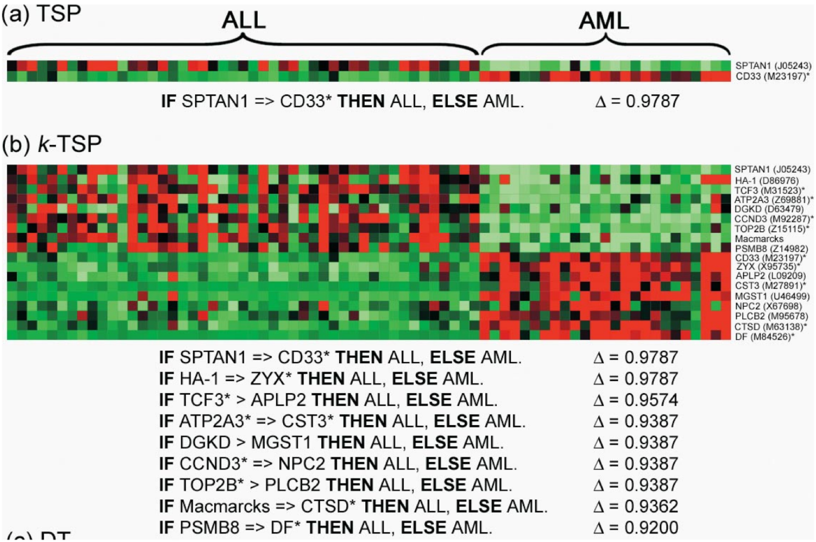
Breast cancer prognosis from gene expression data (survival logistic regression)

Outline

1. SUQUAN: Supervised full quantile normalization (w. Marine Le Morvan)

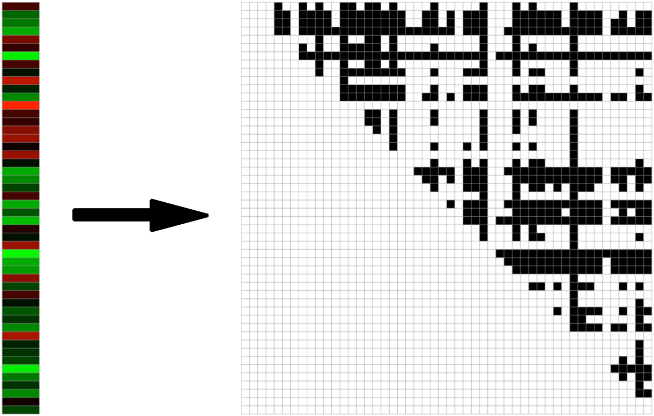
2. Learning from pairwise comparisons with the Kendall and Mallows kernels (w. Yunlong Jiao)

An idea: Top scoring pairs (TSP)



(Geman et al., 2004; Tan et al., 2005; Leek, 2009)

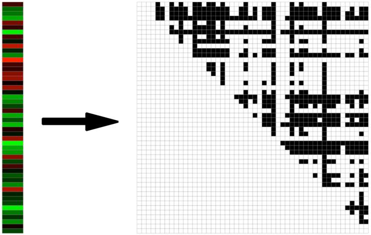
More generally: all pairwise comparisons



One sample x
p features

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

Remark: representation of the symmetric group

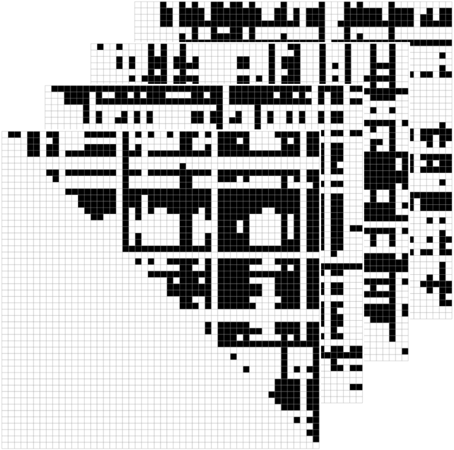


One sample x
p features

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

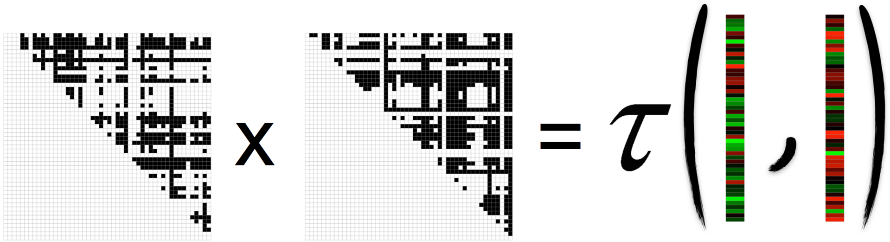
- Obviously, this representation as $O(p^2)$ bits exists for any ranking or permutation of p items
- Many other applications in learning over rankings, learning to rank, learning permutations etc...
- We are interested particularly in practical solutions when p is large

Practical challenge



- Need to store $O(p^2)$ bits per sample
- Need to train a model in $O(p^2)$ dimensions

Kernel trick



O(p^2) O(p log(p))

Good news for SVM and kernel methods!

More formally

- For two permutations σ, σ' let $n_c(\sigma, \sigma')$ (resp. $n_d(\sigma, \sigma')$) the number of concordant (resp. discordant) pairs.
- The Kendall kernel (a.k.a. Kendall tau coefficient) is defined as

$$K_\tau(\sigma, \sigma') = \frac{n_c(\sigma, \sigma') - n_d(\sigma, \sigma')}{\binom{p}{2}}.$$

- The Mallows kernel is defined for any $\lambda \geq 0$ by

$$K_M^\lambda(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')}.$$

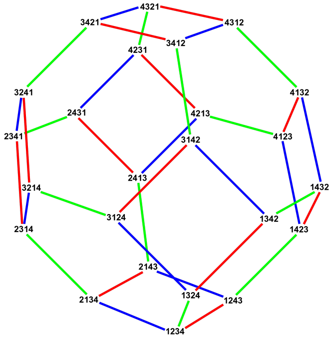
Theorem (Jiao and V., 2015)

The Kendall and Mallows kernels are positive definite.

Theorem (Knight, 1966)

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

Related work



Cayley graph of $\mathbb{S}_4$

- Kondor and Barbarosa (2010) proposed the diffusion kernel on the Cayley graph of the symmetric group generated by adjacent transpositions.
- Computationally intensive ($O(p^p)$)
- Mallows kernel is written as

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

where $n_d(\sigma, \sigma')$ is the shortest path distance on the Cayley graph.
- It can be computed in $O(p \log p)$

Application: supervised classification

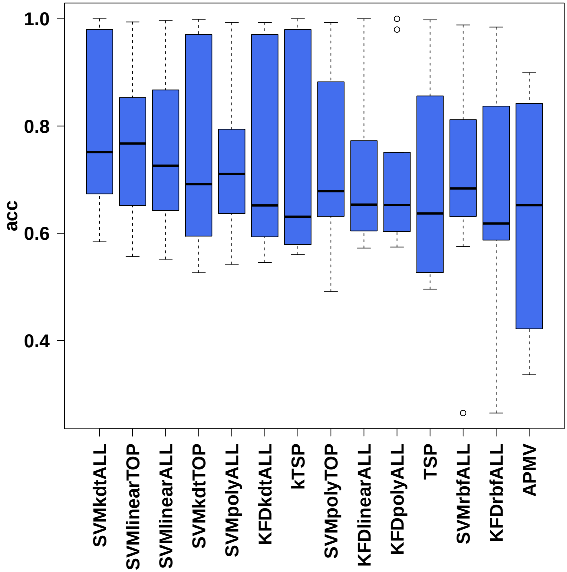
Datasets

Dataset	No. of features	No. of samples (training/test)	
		C_1	C_2
Breast Cancer 1	23624	44/7 (Non-relapse)	32/12 (Relapse)
Breast Cancer 2	22283	142 (Non-relapse)	56 (Relapse)
Breast Cancer 3	22283	71 (Poor Prognosis)	138 (Good Prognosis)
Colon Tumor	2000	40 (Tumor)	22 (Normal)
Lung Cancer 1	7129	24 (Poor Prognosis)	62 (Good Prognosis)
Lung Cancer 2	12533	16/134 (ADCA)	16/15 (MPM)
Medulloblastoma	7129	39 (Failure)	21 (Survivor)
Ovarian Cancer	15154	162 (Cancer)	91 (Normal)
Prostate Cancer 1	12600	50/9 (Normal)	52/25 (Tumor)
Prostate Cancer 2	12600	13 (Non-relapse)	8 (Relapse)

Methods

- Kernel machines Support Vector Machines (SVM) and Kernel Fisher Discriminant (KFD) with Kendall kernel, linear kernel, Gaussian RBF kernel, polynomial kernel.
- Top Scoring Pairs (TSP) classifiers [**?**].
- Hybrid scheme of SVM + TSP feature selection algorithm.

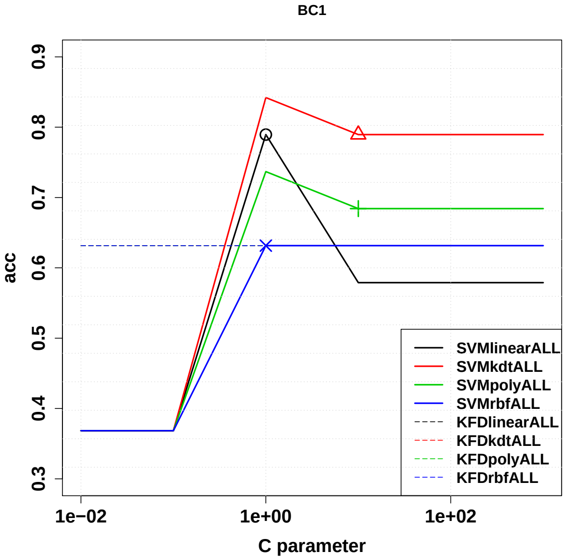
Results



Kendall kernel SVM

- **Competitive accuracy!**
- Less sensitive to regularization parameter!
- No need for feature selection!

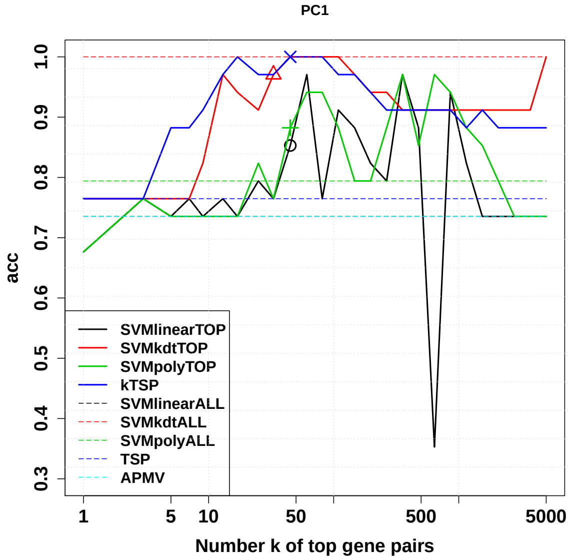
Results



Kendall kernel SVM

- Competitive accuracy!
- **Less sensitive to regularization parameter!**
- No need for feature selection!

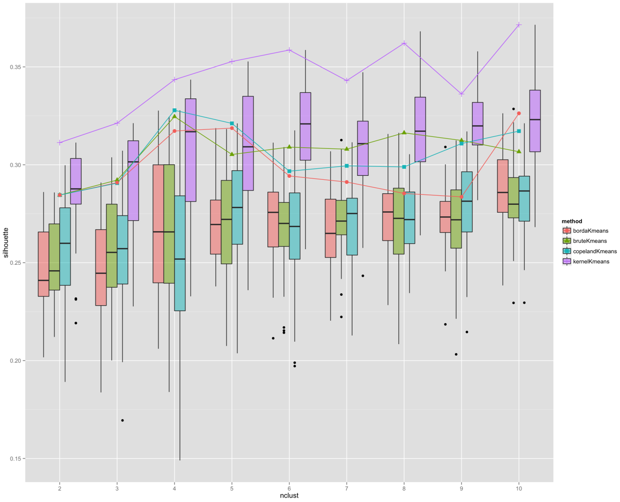
Results



Kendall kernel SVM

- Competitive accuracy!
- Less sensitive to regularization parameter!
- **No need for feature selection!**

Application: clustering



- APA data (full rankings)
- $n = 5738$, $p = 5$
- (new) Kernel k-means vs (standard) k-means in $\mathbb{S}_5$
- Show silhouette as a function of number of clusters (higher better)

Extension to partial rankings

- Two interesting types of partial rankings are interleaving partial ranking

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k}, \quad k \leq n.$$

and top-k partial ranking

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{\mathrm{rest}}, \quad k \leq n.$$

- Partial rankings can be uniquely represented by a set of permutations compatible with all the observed partial orders.

Theorem

For these two particular types of partial rankings, the convolution kernel (Haussler, 1999) induced by Kendall kernel

$$K_\tau^\star(R, R') = \frac{1}{|R||R'|} \sum_{\sigma \in R} \sum_{\sigma' \in R'} K_\tau(\sigma, \sigma')$$

can be evaluated in $O(k \log k)$ time.

Extension to partial rankings

- Two interesting types of partial rankings are interleaving partial ranking

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k}, \quad k \leq n.$$

and top-k partial ranking

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{\mathrm{rest}}, \quad k \leq n.$$

- Partial rankings can be uniquely represented by a set of permutations compatible with all the observed partial orders.

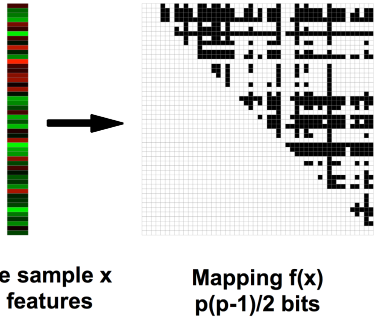
Theorem

For these two particular types of partial rankings, the convolution kernel (Haussler, 1999) induced by Kendall kernel

$$K_\tau^\star(R, R') = \frac{1}{|R||R'|} \sum_{\sigma \in R} \sum_{\sigma' \in R'} K_\tau(\sigma, \sigma')$$

can be evaluated in $O(k \log k)$ *time.*

Extension to smoother, continuous representations

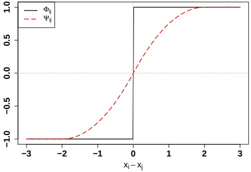


- Instead of $\Phi : \mathbb{R}^p \to \{0,1\}^{p(p-1)/2}$, consider the continuous mapping $\Psi_a : \mathbb{R}^p \to \mathbb{R}^{p(p-1)/2}$:

$$\Psi_a(x) = \mathbb{E}\Phi(x+\epsilon) \quad \text{with} \quad \epsilon \sim (\mathcal{U}[-\frac{a}{2},\frac{a}{2}])^n$$

- Corresponding kernel $G_a(x,x') = \Psi_a(x)^\top \Psi_a(x')$

Computation of $G(x, x')$



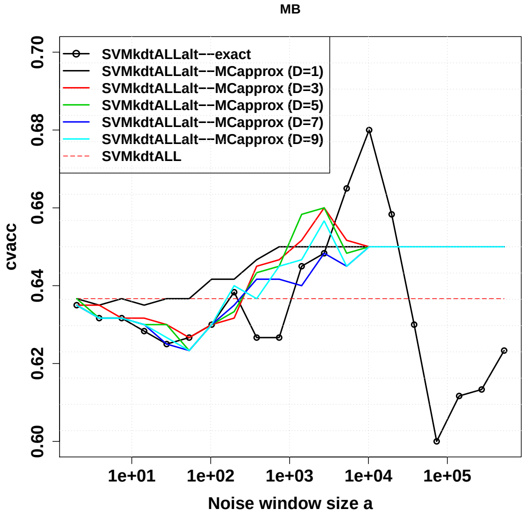
- $G_a(x, x')$ can be computed exactly in $O(p^2)$ by explicit computation of $\Psi_a(x)$ in $\mathbb{R}^{p(p-1)/2}$
- $G_a(x, x')$ can be computed approximately in $O(D^2 p \log p)$ by Monte-Carlo approximation:

$$\tilde{G}_a(x, x') = \frac{1}{D^2} \sum_{i,j=1}^{D} K(x + \epsilon_i, x' + \epsilon'_j)$$

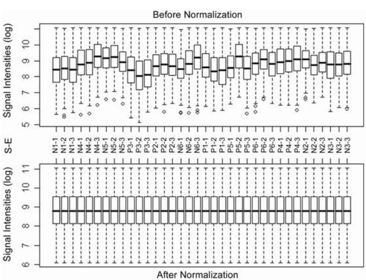
- Theorem: for supervised learning, Monte-Carlo approximation is better[1] than exact computation when $n = o(p^{1/3})$

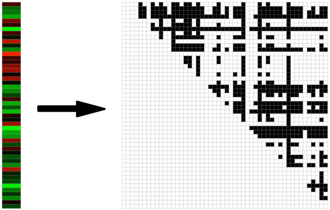
[1] faster for the same accuracy

Performance of $G_a(x, x)$



Conclusion





One sample x
p features

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

- Full quantile normalization as matrix learning
- A representation of vectors that only depends on the relative order of features
- A tractable $O(p \log p)$ kernel for (partial) ranking and permutations
- Open questions
  - higher-order comparisons
  - primal approximation in less than $O(p^2)$ dimension
  - learning the representation

Thanks

erc

SEVENTH FRAMEWORK PROGRAMME

INVESTISSEMENTS D'AVENIR

AGENCE NATIONALE DE LA RECHERCHE ANR

BIOMÉRIEUX

MINES ParisTech

institutCurie
Together, let's beat cancer.

Inserm
Institut national
de la santé et de la recherche médicale

PSL
RESEARCH UNIVERSITY

The University of California Berkeley 1868

The Adolph C. and Mary Sprague
Miller Institute for Basic
Research in Science
University of California, Berkeley

SIMONS INSTITUTE
for the Theory of Computing