

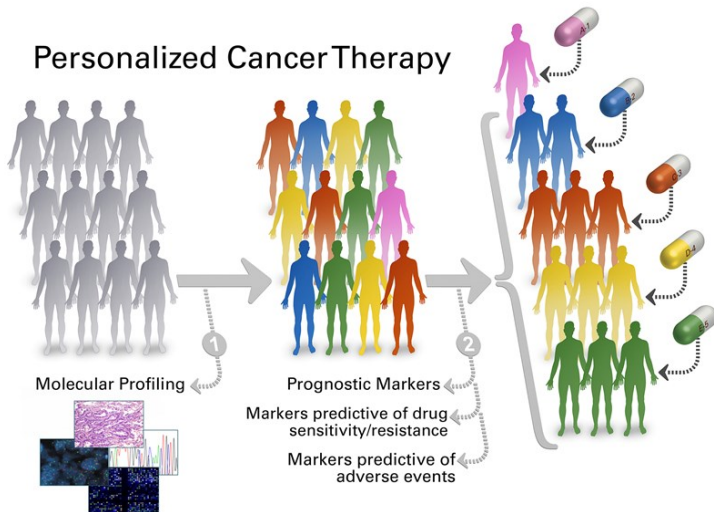
Machine learning for patient stratification from genomic data

Jean-Philippe Vert



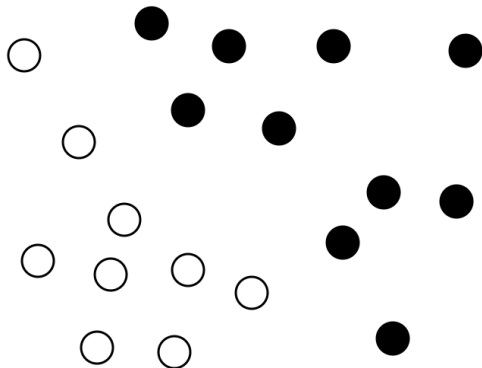
CBI colloquium, University of Saarland, January 18, 2017

Personalized Cancer Therapy



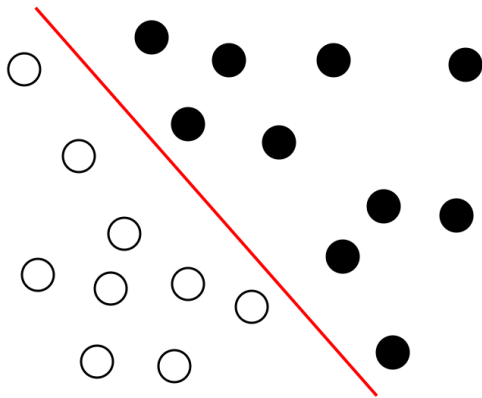
Mathematical model

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



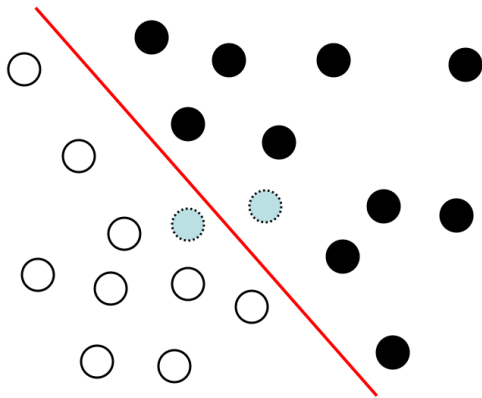
Mathematical model

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



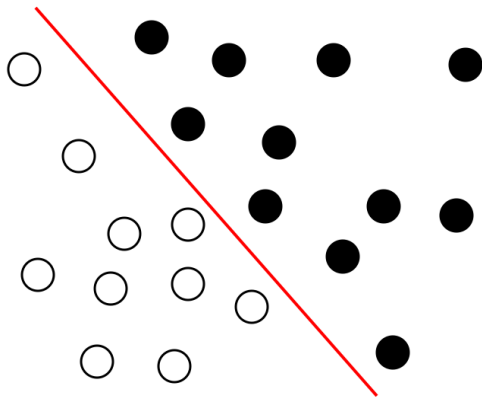
Mathematical model

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



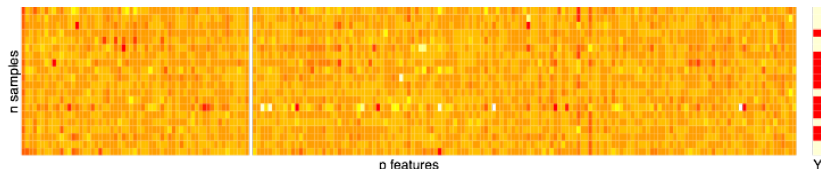
Mathematical model

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

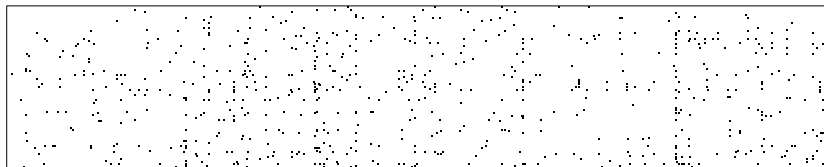


Real data: $n \lll p$

- Gene expression



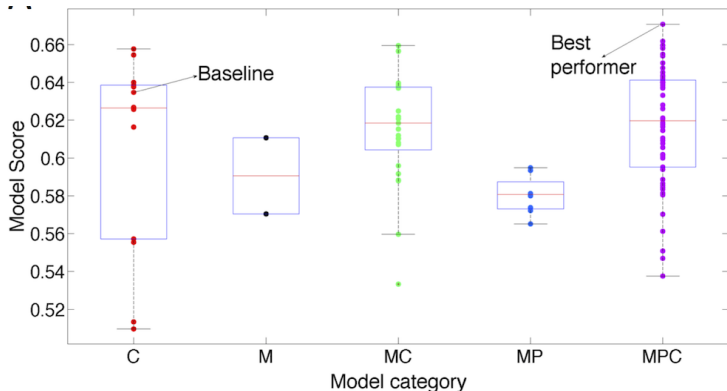
- Somatic mutations



- $n = 10^2 \sim 10^4$ (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)
- P: incorporate prior knowledge

Consequence: unstable biomarker selection

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

* 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, 12040 115th Avenue NE, Kirkland, Washington 98034

70 genes (Nature, 2002)

Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer

Yixin Wang, Jan G M Kljin, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

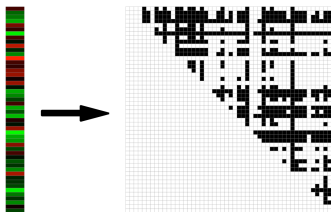
76 genes (Lancet, 2005)

3 genes in common

van 't Veer et al. (2002); Wang et al. (2005)

Some research directions

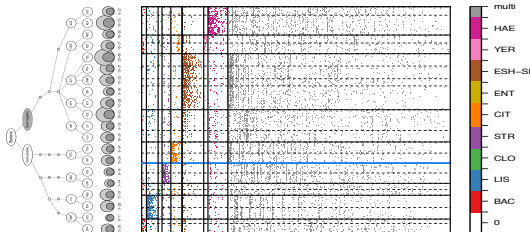
- Find a better **representation**



One sample \times
 p features

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

- Incorporate **prior knowledge**



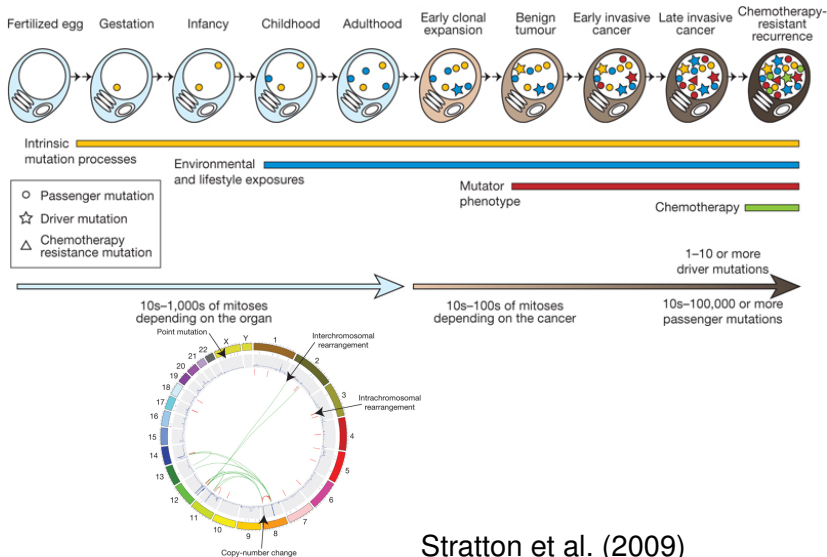
Outline

- 1 Learning from mutation data
- 2 Supervised quantile normalization
- 3 The Kendall and Mallows kernels
- 4 Conclusion

Outline

- 1 Learning from mutation data
- 2 Supervised quantile normalization
- 3 The Kendall and Mallows kernels
- 4 Conclusion

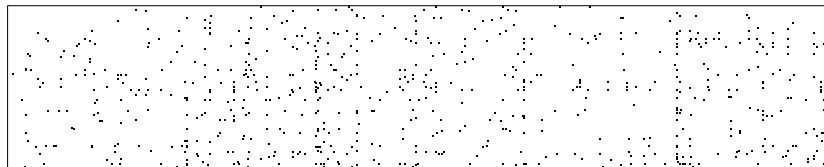
Somatic mutations in cancer



Stratton et al. (2009)

Large-scale efforts to collect somatic mutations

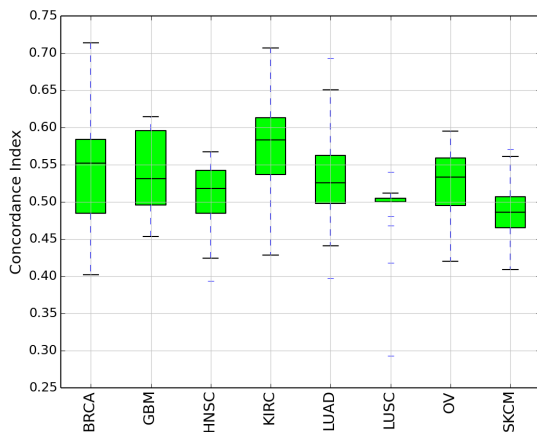
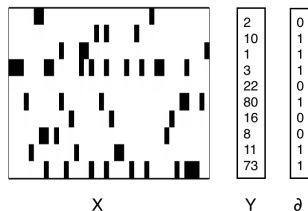
- 3,378 samples with survival information from 8 cancer types
- downloaded from the TCGA / cBioPortal portals.



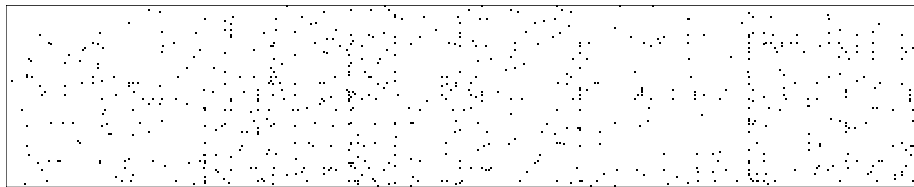
Cancer type	Patients	Genes
LUAD (Lung adenocarcinoma)	430	20 596
SKCM (Skin cutaneous melanoma)	307	17 463
GBM (Glioblastoma multiforme)	265	14 750
BRCA (Breast invasive carcinoma)	945	16 806
KIRC (Kidney renal clear cell carcinoma)	411	10 609
HNSC (Head and Neck squamous cell carcinoma)	388	17 022
LUSC (Lung squamous cell carcinoma)	169	13 590
OV (Ovarian serous cystadenocarcinoma)	363	10 195

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



Changing the representation?

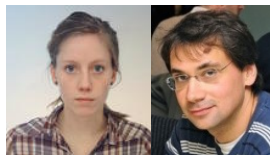


Can we replace

$x \in \{0, 1\}^p$ with p very large, very sparse

by a representation with more information shared between samples

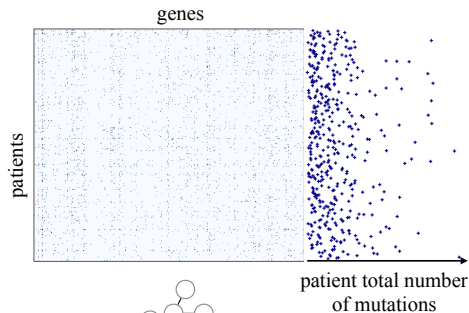
$\Phi(x) \in \mathcal{H}$?



NetNorm Overview (Le Morvan et al., 2016)

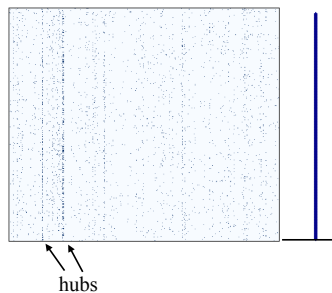
- **Modify** the binary vector $x \in \{0, 1\}^p$ of each patient by **adding or removing mutations**, using a **gene network** as prior knowledge
- After Netnorm, all patients $\Phi(x) \in \{0, 1\}^p$ have the **same number of (pseudo-)mutations**

Raw binary mutation matrix



Gene-gene interaction network

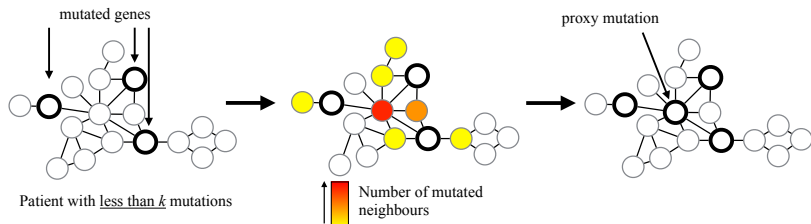
NetNorM binary mutation matrix



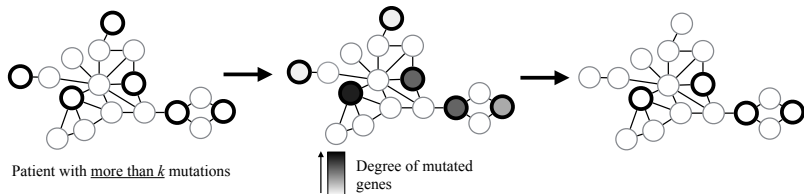
hubs

NetNorm detail ($k=4$)

- 1 **Add** mutations for patients with **few** (less than k) mutations



- 2 **Remove** mutations for patients for **many** (more than k) mutations



Network-based stratification of tumor mutations

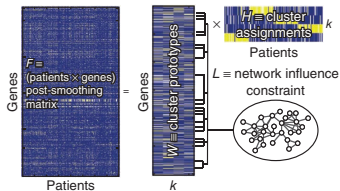
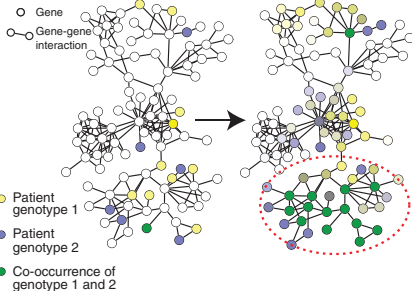
Matan Hofree¹, John P Shen², Hannah Carter², Andrew Gross³ & Trey Ideker¹⁻³

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).

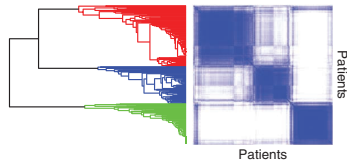
RECEIVED 14 FEBRUARY; ACCEPTED 12 AUGUST; PUBLISHED ONLINE 15 SEPTEMBER 2013; DOI:10.1038/NMETH.2651

1108 | VOL.10 NO.11 | NOVEMBER 2013 | NATURE METHODS

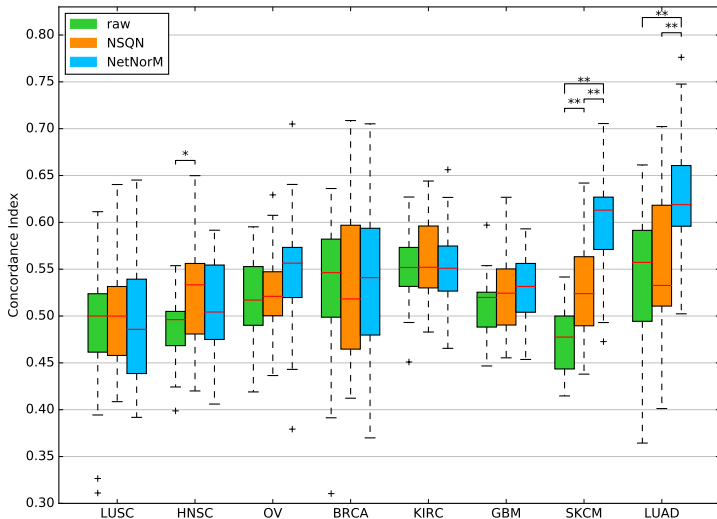
Network smoothing:



d Network-based stratification



Performance on survival prediction

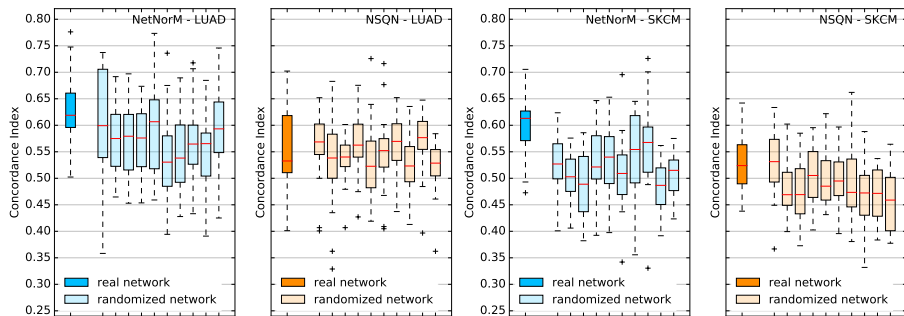


Use Pathway Commons as gene network.

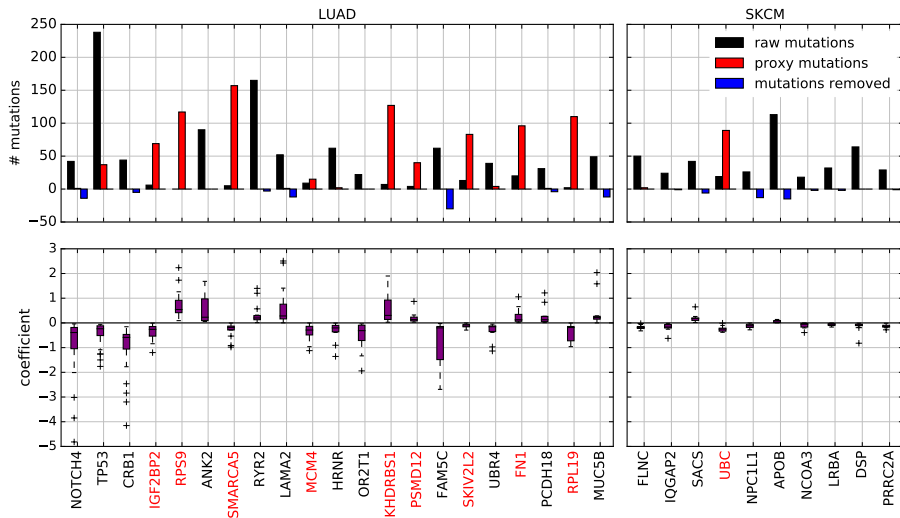
NSQN = Network Smoothing / Quantile Normalization (Hofree et al., 2013)

NetNorM and NSQN benefit from biological information in the gene network

Comparison with 10 randomly permuted networks:

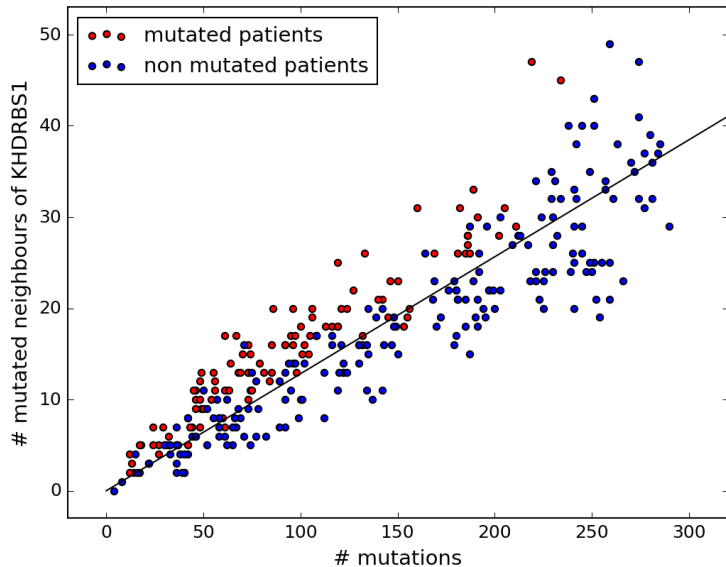


Selected genes represent "true" or "proxy" mutations

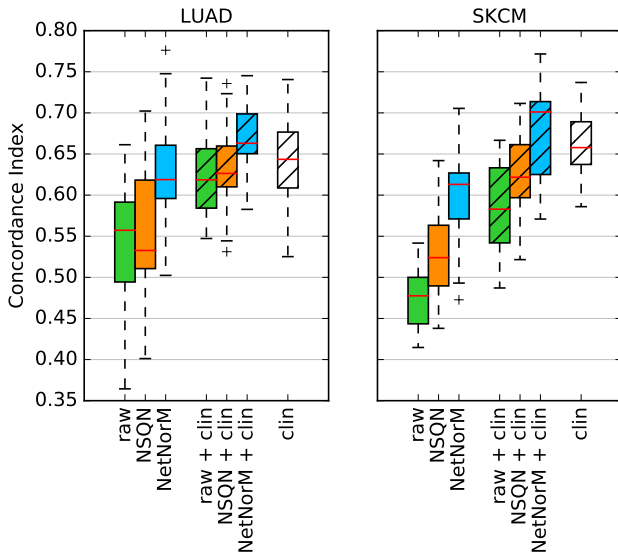


Genes selected in at least 50% of the cross-validated sparse SVM model

Proxy mutations encode local mutational burden



Adding good old clinical factors



Combination by averaging predictions

Outline

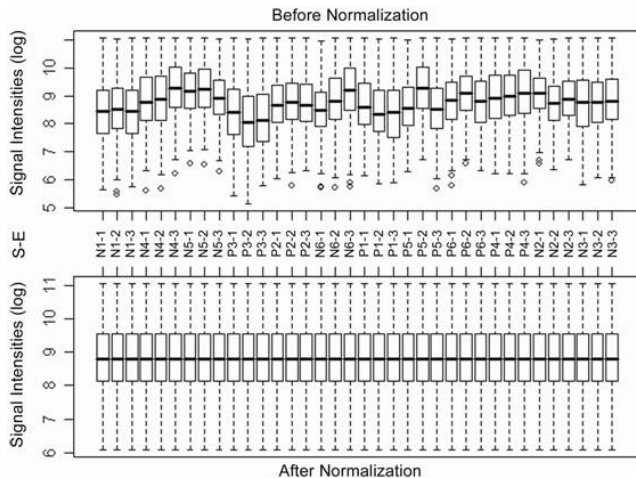
- 1 Learning from mutation data
- 2 Supervised quantile normalization**
- 3 The Kendall and Mallows kernels
- 4 Conclusion

Joint work with



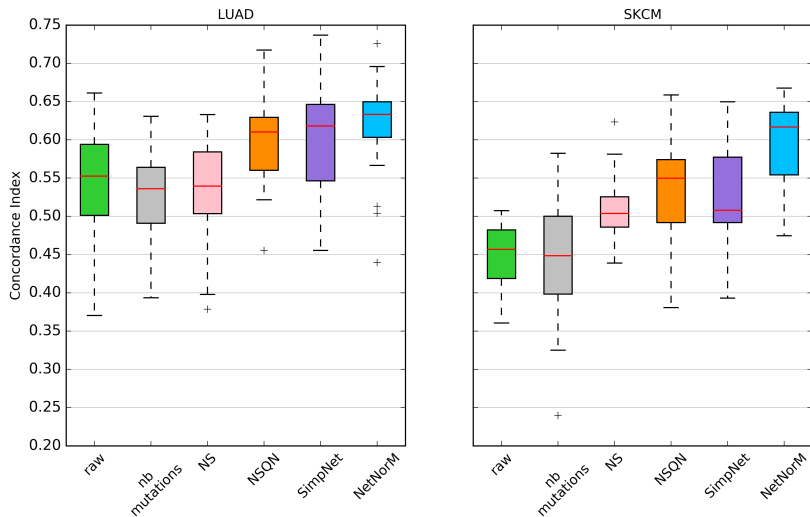
Marine Le Morvan

Standard full quantile normalization



Typically followed by a predictive model $f(X)$ on the normalized data

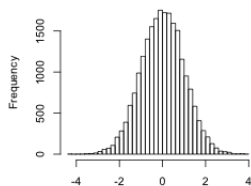
Choosing a "good" target distributions is important



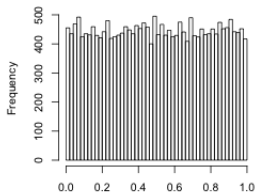
Cancer prognosis from somatic mutations

How to choose a "good" target distribution?

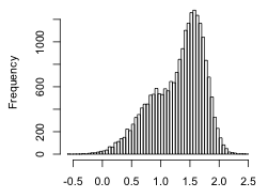
gaussian distribution (mean=0, sd=1)



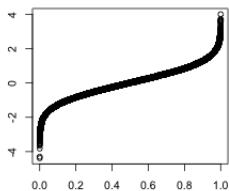
uniform distribution



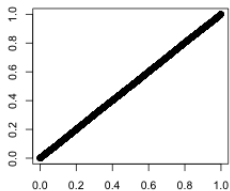
bigaussian distribution



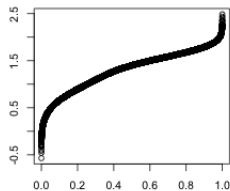
quantile function (->gaussian)



quantile function (-> uniform)



quantile function (->bigaussian)



Learning the target distribution

- x_1, \dots, x_n a set of p -dimensional samples
- $f \in \mathbb{R}^p$ a non-decreasing target distribution (CDF)
- For $x \in \mathbb{R}^p$, let $\Phi_f(x) \in \mathbb{R}^p$ be the data after QN with target distribution f
- **Standard approaches** (NSQN, NetNorM, ...)
 - 1 Fix f arbitrarily
 - 2 QN all samples to get $\Phi_f(x_1), \dots, \Phi_f(x_n)$
 - 3 Learn a generalized linear model (w, b) on normalized data:

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

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

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

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

$$x = \begin{pmatrix} 4.5 \\ 1.2 \\ 10.1 \\ 8.9 \end{pmatrix} \quad \Pi_x = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad f = \begin{pmatrix} 0 \\ 1 \\ 3 \\ 4 \end{pmatrix}$$

- Quantile normalized x with target distribution f is:

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

- SUQUAN solves

$$\begin{aligned} \min_{w,b,f} \frac{1}{n} \sum_{i=1}^n \ell \left(w^\top \Pi_{x_i} f + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \\ = \min_{w,b,f} \frac{1}{n} \sum_{i=1}^n \ell \left(\langle w f^\top, \Pi_{x_i} \rangle + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \end{aligned} \tag{1}$$

- A particular **rank-1 matrix optimization**, x is **replaced by Π_x**
- Solved by alternatively optimizing f and w

Experiments

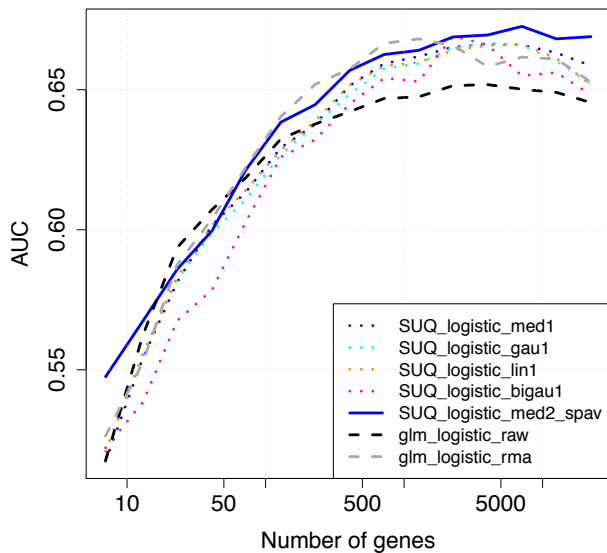
$$\min_{w,b,f} \frac{1}{n} \sum_{i=1}^n \ell_i \left(w^\top \Phi_f(x_i) + b \right) + \frac{\lambda}{2} \|w\|_2^2 + \frac{\gamma}{2} \sum_{j=1}^{p-1} (f_{j+1} - f_j)^2$$

- Breast cancer prognosis from gene expression data.
- Two classes of patients: those who relapsed within 6 years of diagnosis and those who did not.

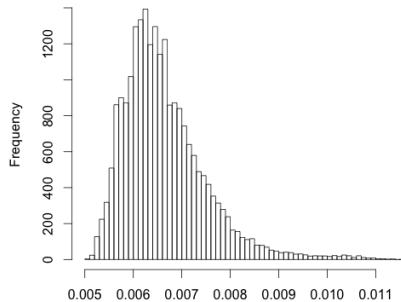
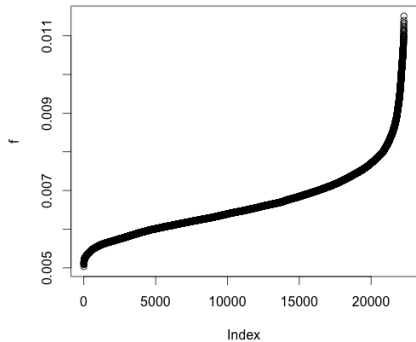
Dataset name	# genes	# patients	# positives	% positives
GSE7390	22283	189	58	0.31
GSE4922	22283	225	73	0.32
GSE2990	22283	106	32	0.30
GSE2034	22283	271	104	0.38
GSE1456	22283	141	37	0.26

Performance

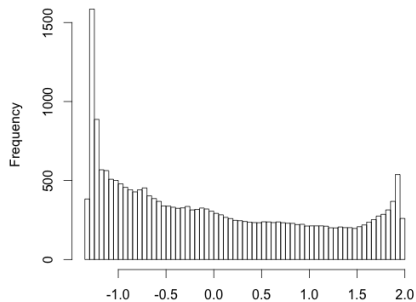
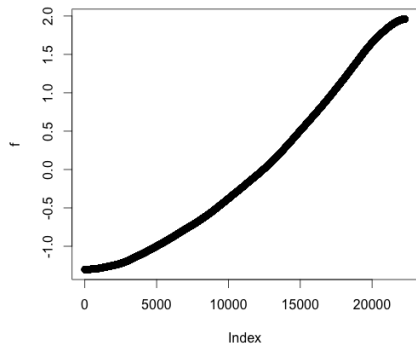
average over all datasets



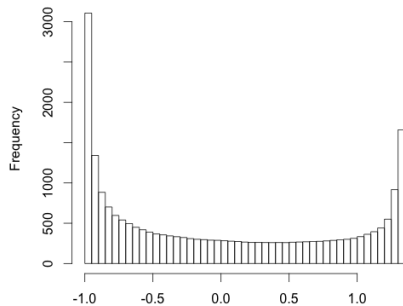
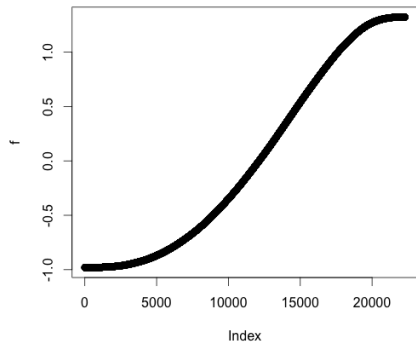
Estimated distribution: iteration=0



Estimated distribution: iteration=1



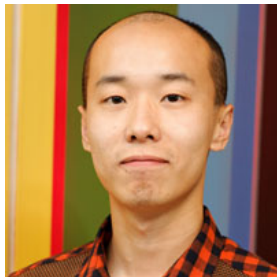
Estimated distribution: iteration=2



Outline

- 1 Learning from mutation data
- 2 Supervised quantile normalization
- 3 The Kendall and Mallows kernels**
- 4 Conclusion

Joint work with

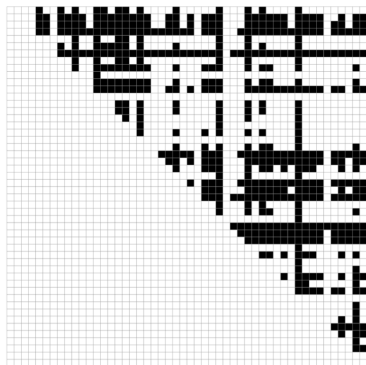


Yunlong Jiao

An idea: all pairwise comparisons

Replace $x \in \mathbb{R}^p$ by $\Phi(x) \in \{0, 1\}^{p(p-1)/2}$:

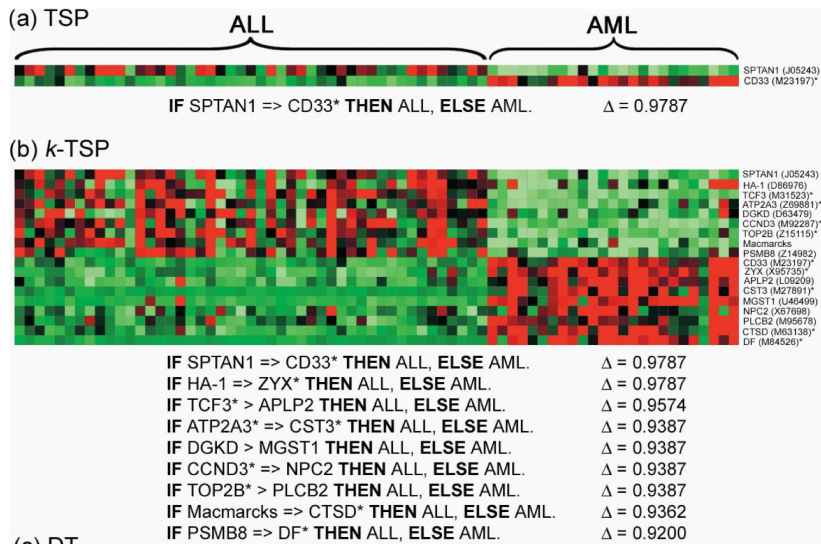
$$\Phi_{i,j}(x) = \begin{cases} 1 & \text{if } x_i \leq x_j, \\ 0 & \text{otherwise.} \end{cases}$$



**One sample x
 p features**

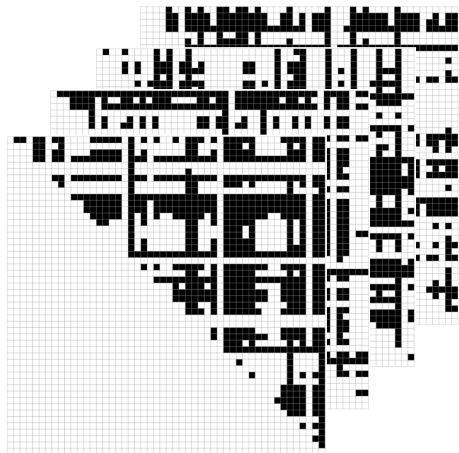
**Mapping $f(x)$
 $p(p-1)/2$ bits**

Related work: Top scoring pairs (TSP)



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

Practical challenge



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

Theorem (Wahba, Schölkopf, ...)

Training a linear model over a representation $\Phi(x) \in \mathbb{R}^Q$ of the form:

$$\min_{w \in \mathbb{R}^Q} \frac{1}{n} \sum_{i=1}^n \ell(w^\top \Phi(x_i), y_i) + \lambda \|w\|^2$$

can be done efficiently, independently of Q , if the kernel

$$K(x, x') = \Phi(x)^\top \Phi(x')$$

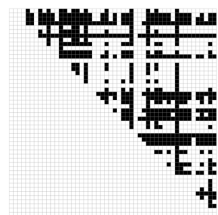
can be computed efficiently.

Ex: ridge regression, $O(Q^3 + nQ^2)$ becomes $O(n^3 + n^2 T)$

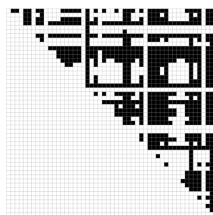
Other: SVM, logistic regression, Cox model, survival SVM, ...

Kernel trick for us: Kendall's τ

$$\Phi(x)^\top \Phi(x') = \tau(x, x') \quad (\text{up to a scaling})$$



\times



$$= \tau \left(\begin{array}{c} \text{red} \\ \text{green} \\ \text{red} \\ \text{green} \\ \text{red} \\ \text{green} \\ \text{red} \\ \text{green} \\ \text{red} \\ \text{green} \end{array} , \begin{array}{c} \text{red} \\ \text{green} \\ \text{red} \\ \text{green} \\ \text{red} \\ \text{green} \\ \text{red} \\ \text{green} \\ \text{red} \\ \text{green} \end{array} \right)$$

$O(p^2)$

$O(p \log(p))$

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

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

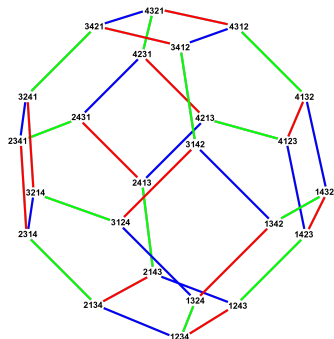
Theorem (Jiao and V., 2015)

*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 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

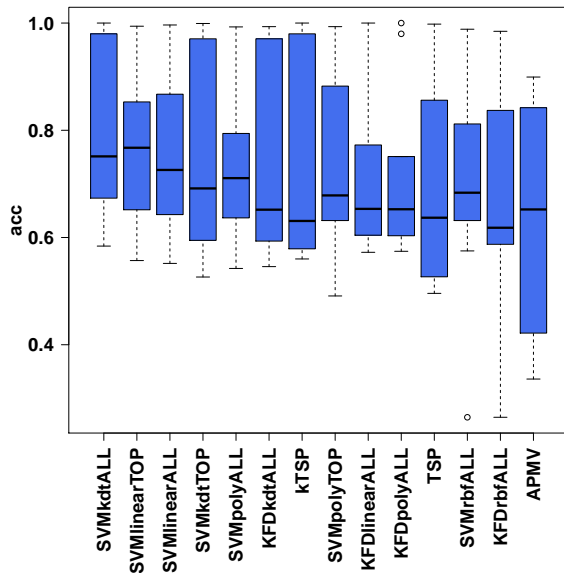
Datasets

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 Tan et al. (2005).
- Hybrid scheme of SVM + TSP feature selection algorithm.

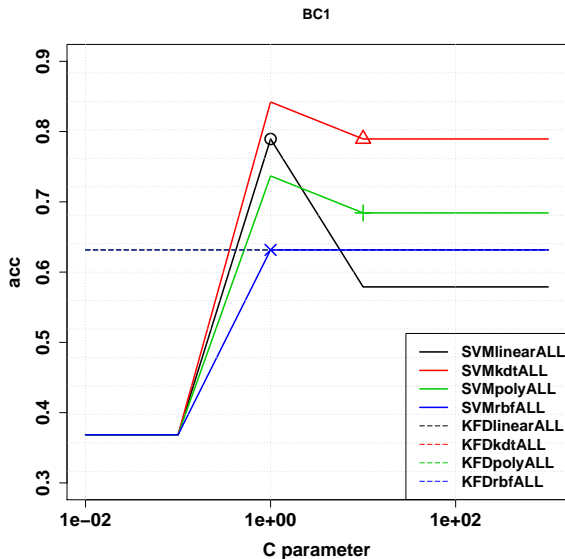
Results



Kendall kernel SVM

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

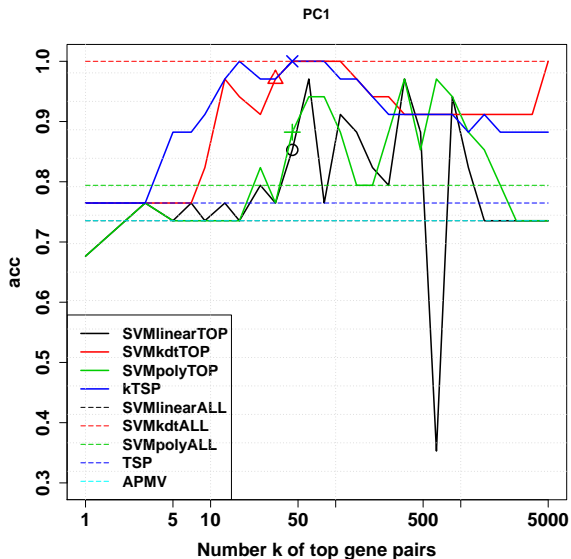
Results



Kendall kernel SVM

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

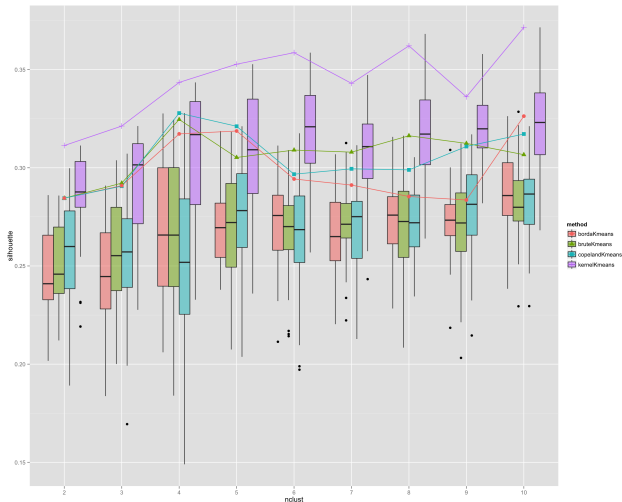
Results



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**

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

and **top-k partial ranking**

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

- 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

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

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

Extension to partial rankings

- Two interesting types of partial rankings are **interleaving partial ranking**

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

and **top-k partial ranking**

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

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

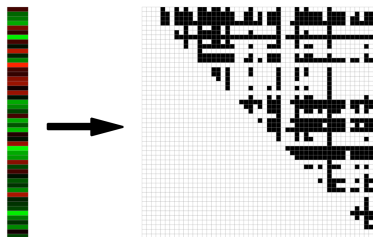
Theorem

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

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

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

Extension to smoother, continuous representations



One sample x
 p features

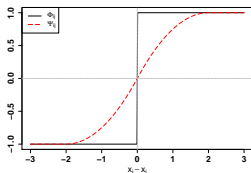
Mapping $f(x)$
 $p(p-1)/2$ bits

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

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

- 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 $\mathbb{R}^{p(p-1)/2}$

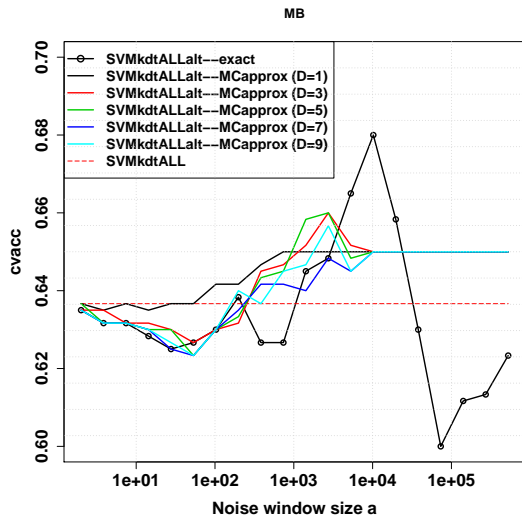
- $G_a(x, x')$ can be computed **approximately** in $O(D^2 p \log p)$ by Monte-Carlo approximation:

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

- Theorem: for supervised learning, Monte-Carlo approximation is better¹ than exact computation when $n = o(p^{1/3})$

¹ faster for the same accuracy

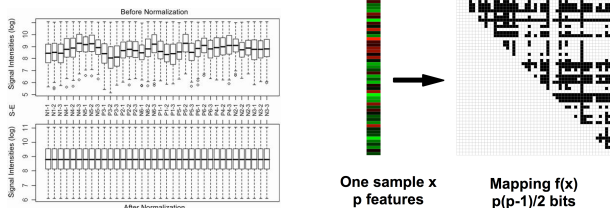
Performance of $G_a(x, x)$



Outline

- 1 Learning from mutation data
- 2 Supervised quantile normalization
- 3 The Kendall and Mallows kernels
- 4 Conclusion**

Conclusion



- Many learning problem in precision medicine are **hard**, machine learning is no magic bullet with $n \ll p$ and complex data
- Understanding the **benefits and cost** of different representations remains very heuristic and sometimes counterintuitive
- NetNorm is one way to use prior knowledge; why it "works" is not fully understood
- Representing omics data as **permutations** has some potential; the information lost about the gene expression values seems irrelevant (SUQUAN, Kendall and Mallow's kernels)
- **Learning representation** is worth investigating

Thanks



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



SIMONS
INSTITUTE
for the Theory of Computing

References

- M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker. Network-based stratification of tumor mutations. *Nat Methods*, 10(11):1108–1115, Nov 2013. doi: 10.1038/nmeth.2651. URL <http://dx.doi.org/10.1038/nmeth.2651>.
- M. Le Morvan, A. Zinovyev, and J.-P. Vert. Netnorm: capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. Technical Report 01341856, HAL, 2016. URL <http://hal.archives-ouvertes.fr/hal-01341856>.
- M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239):719–724, Apr 2009. doi: 10.1038/nature07943. URL <http://dx.doi.org/10.1038/nature07943>.
- A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, and D. Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, Oct 2005. doi: 10.1093/bioinformatics/bti631. URL <http://dx.doi.org/10.1093/bioinformatics/bti631>.
- L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancers. *Nature*, 415(6871):530–536, Jan 2002. doi: 10.1038/415530a. URL <http://dx.doi.org/10.1038/415530a>.

References (cont.)

- Y. Wang, J. Klijn, Y. Zhang, A. Sieuwerts, M. Look, F. Yang, D. Talantov, M. Timmermans, M. Meijer-van Gelder, J. Yu, T. Jatko, E. Berns, D. Atkins, and J. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancers. *Lancet*, 365(9460):671–679, 2005. doi: 10.1016/S0140-6736(05)17947-1. URL [http://dx.doi.org/10.1016/S0140-6736\(05\)17947-1](http://dx.doi.org/10.1016/S0140-6736(05)17947-1).