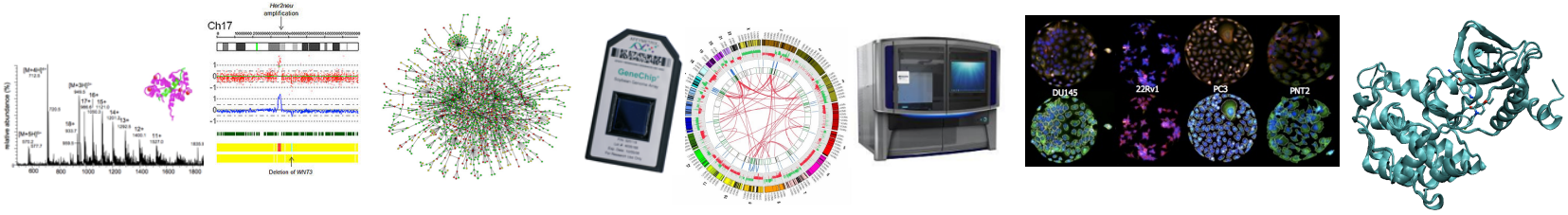
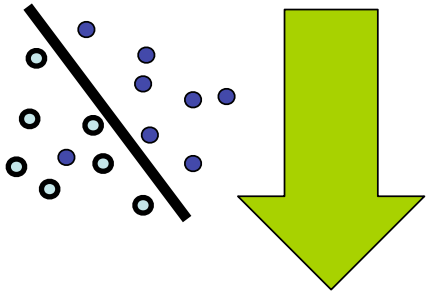
From DNA mutations to embeddings of permutations

Jean-Philippe Vert

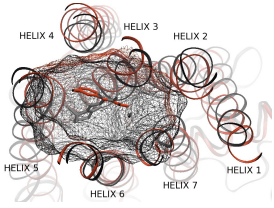
Google / MINES ParisTech

Overview

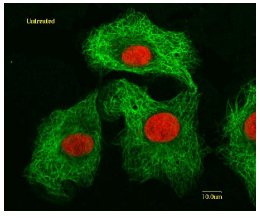




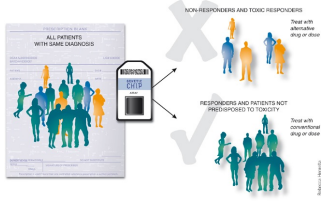
Machine learning
Learning with complex data
Regularization
Scalable algorithms



Molecules
(Epi)-Genomics
Systems biology
Drug design

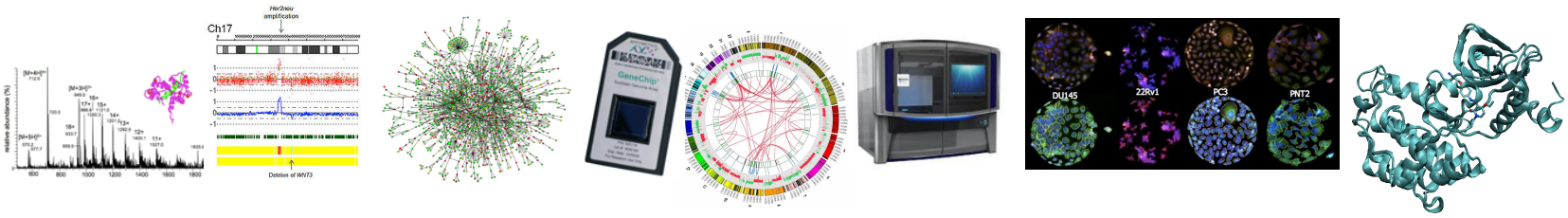


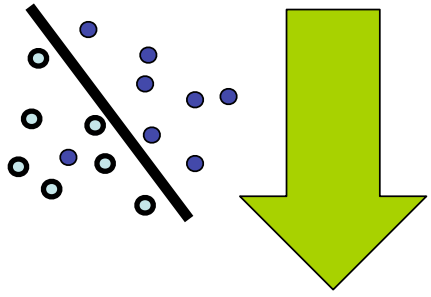
Cells
High-content screening
Single-cell genomics
Tumour heterogeneity



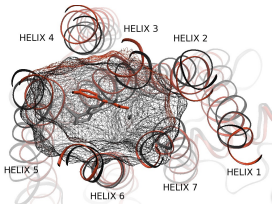
People
Precision medicine
GWAS
Patient monitoring

Overview

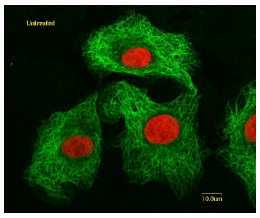




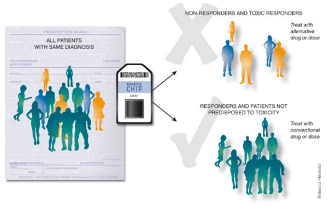
Machine learning
Learning with complex data
Regularization
Scalable algorithms



Molecules
(Epi)-Genomics
Systems biology
Drug design



Cells
High-content screening
Single-cell genomics
Tumour heterogeneity

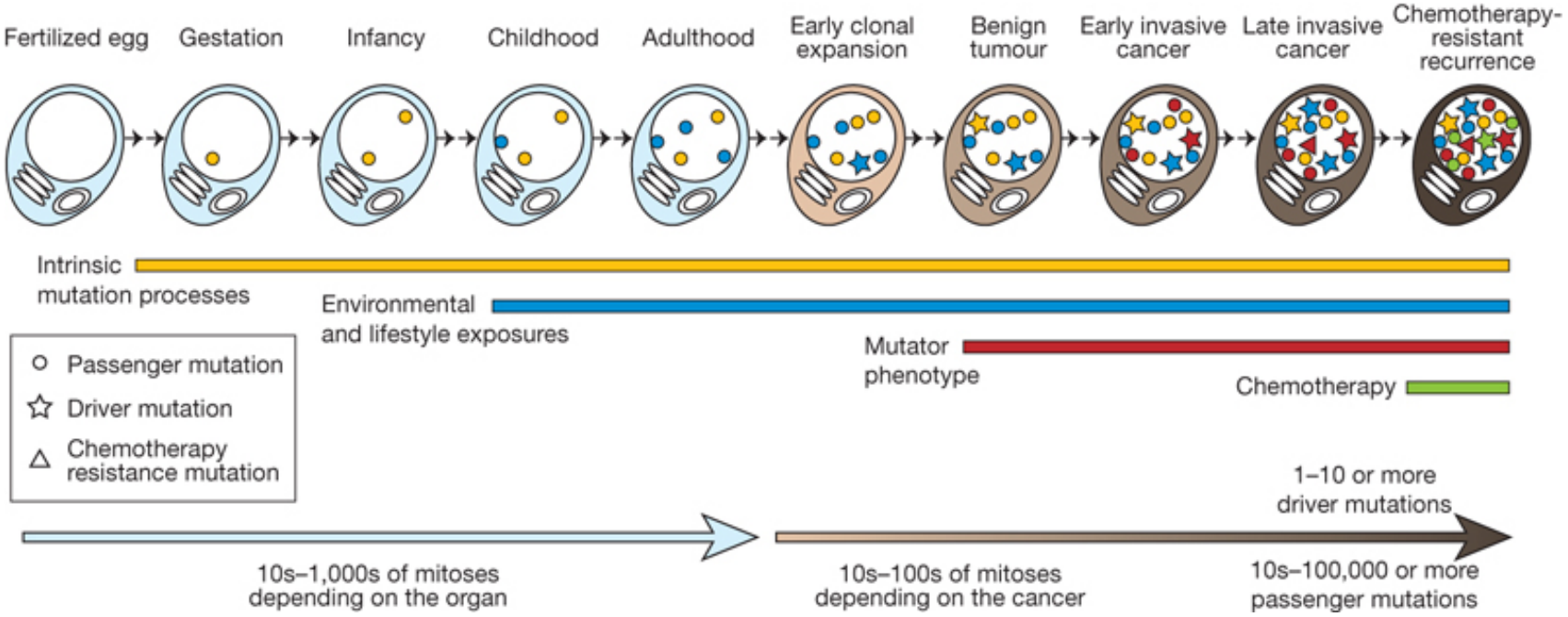


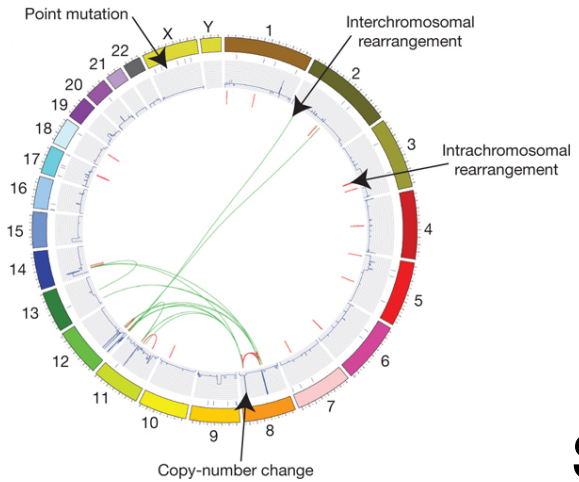
People
Precision medicine
GWAS
Patient monitoring

Outline

Outline

Somatic mutations in cancer

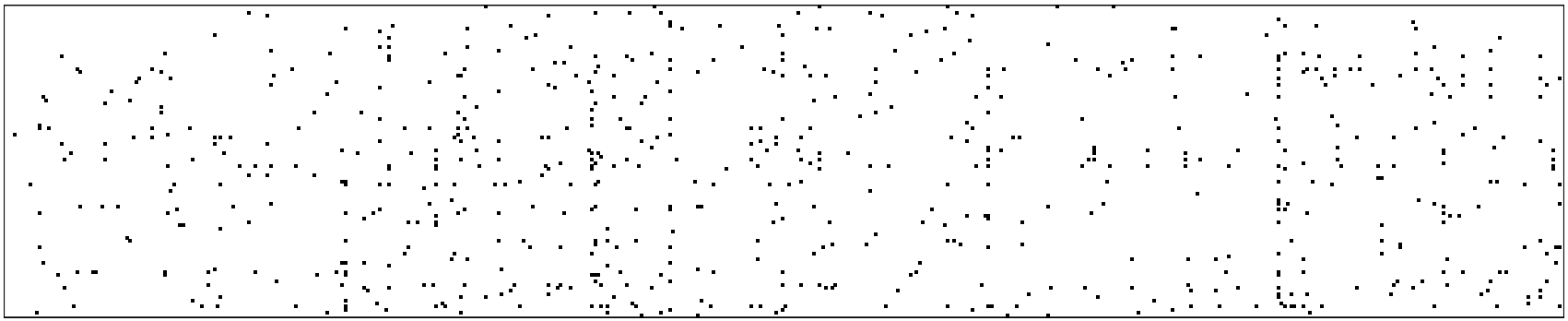




Stratton et al. (2009)

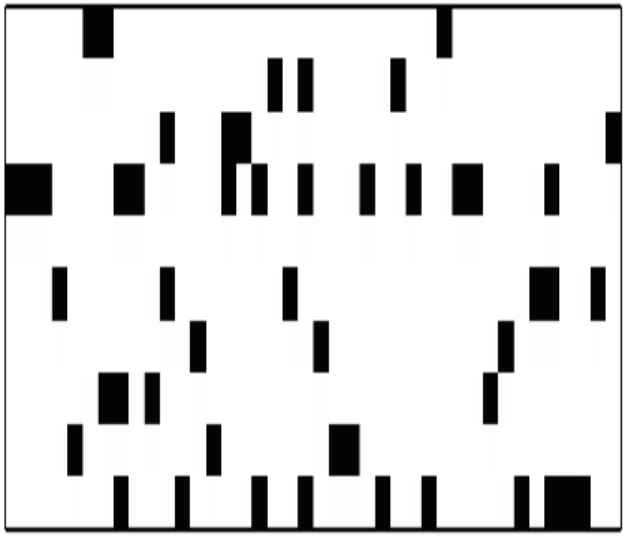
Large-scale efforts to collect somatic mutations

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



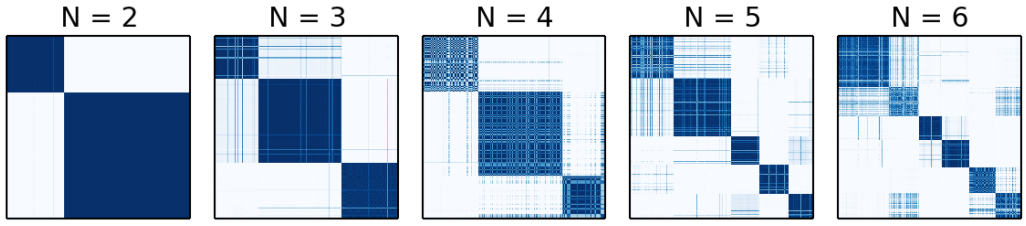
Cancer type	Patients	Genes
LUAD (Lung adenocarcinoma)	430	20 596
SKCM (Skin cutaneous melanoma)	307	17 463
GBM (Glioblastoma multiforme)	265	14 750
BRCA (Breast invasive carcinoma)	945	16 806
KIRC (Kidney renal clear cell carcinoma)	411	10 609
HNSC (Head and Neck squamous cell carcinoma)	388	17 022
LUSC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

Patient stratification (unsupervised) from raw mutation profiles

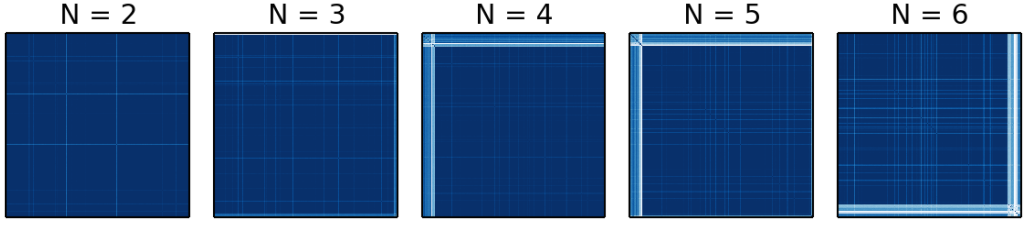


✓ Non-Negative matrix factorisation (NMF)

✓ Desired behaviour:



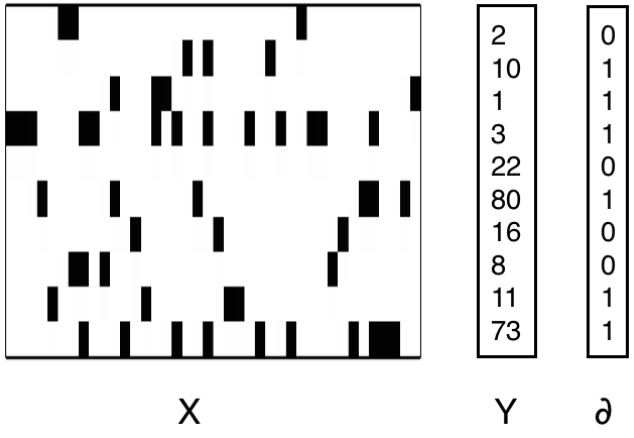
✓ Observed behaviour:

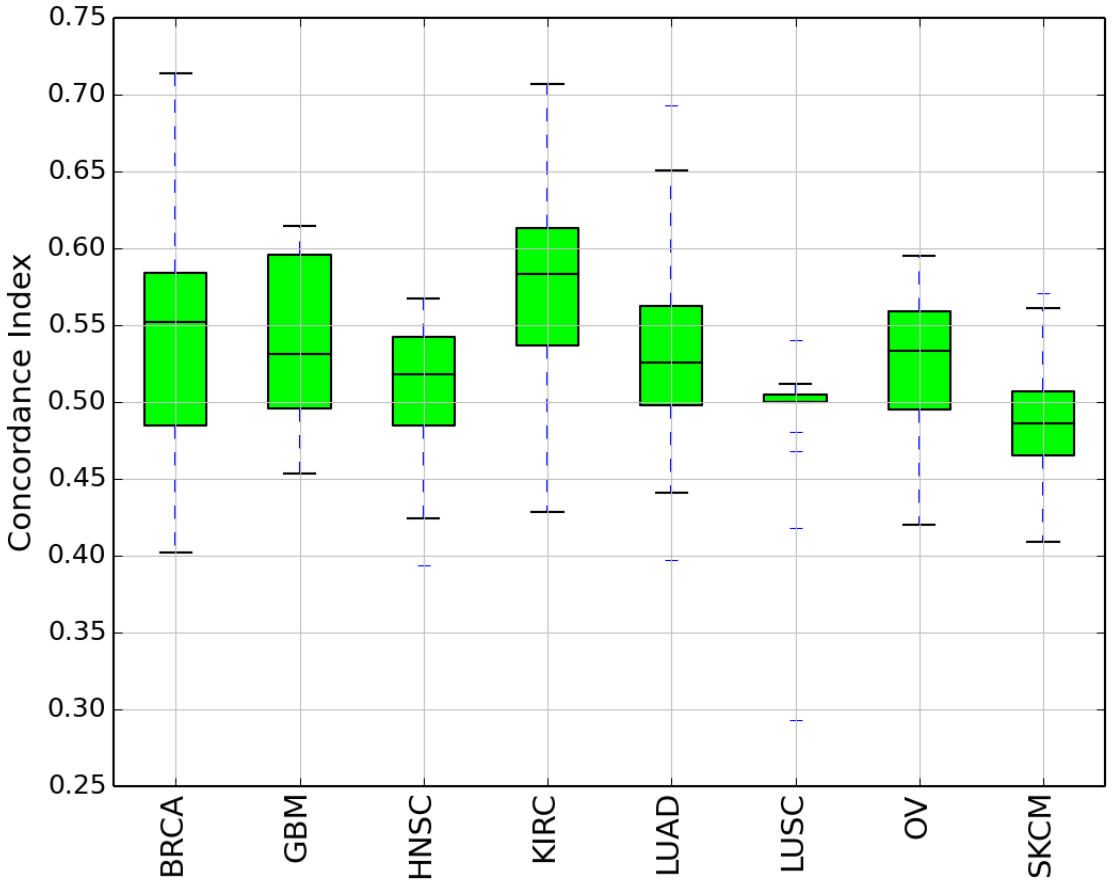


Patients share very few mutated genes!

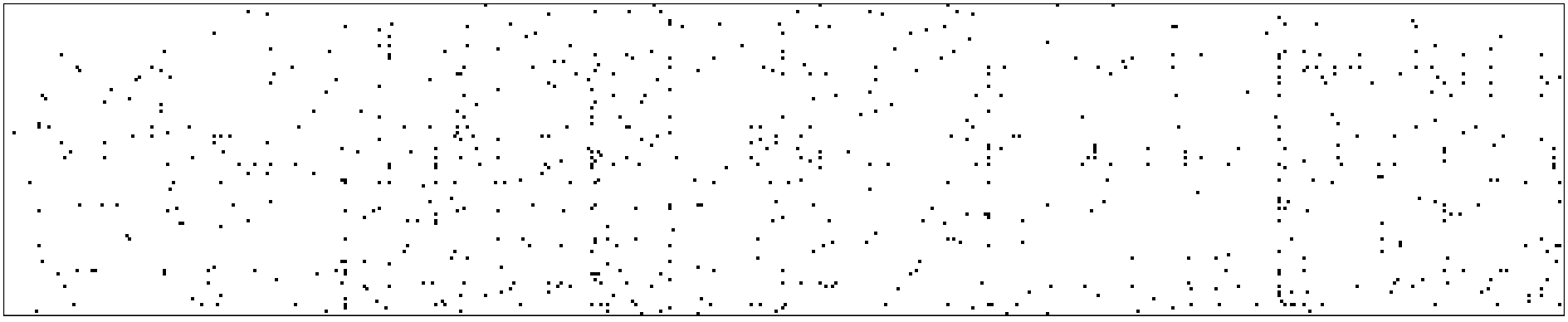
Survival prediction from raw mutation profiles

- 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





Approach: change representation?



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?

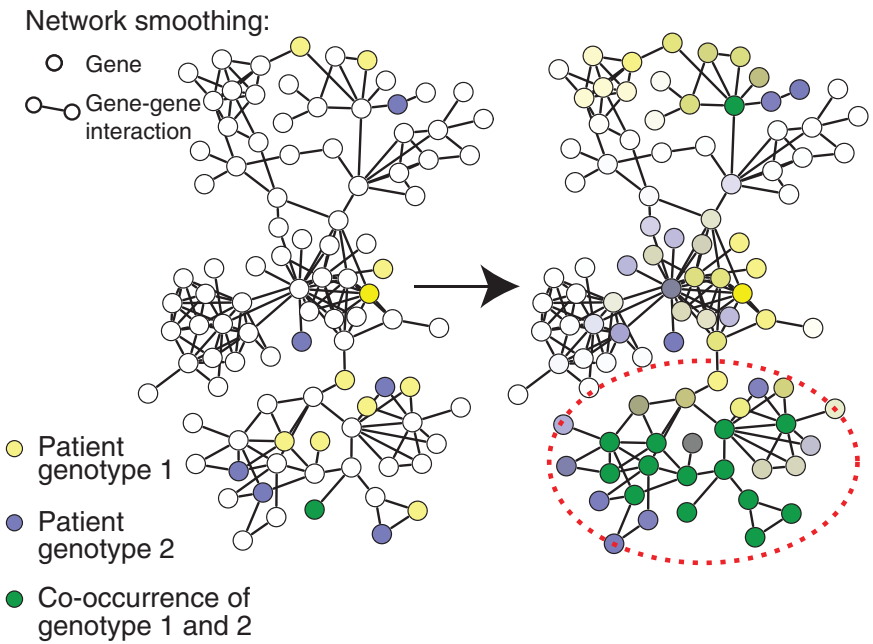
Related work (Hofree et al., 2013)

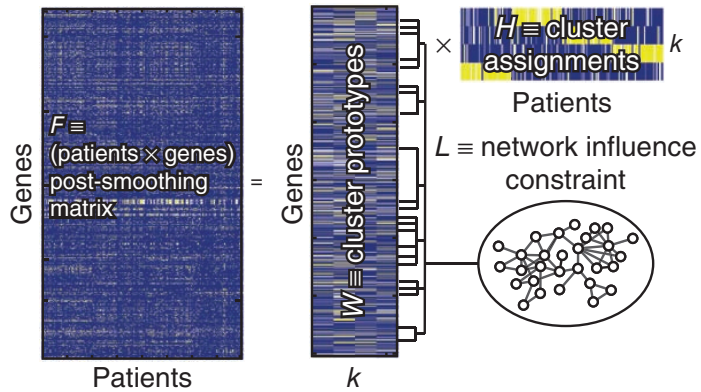
Network-based stratification of tumor mutations

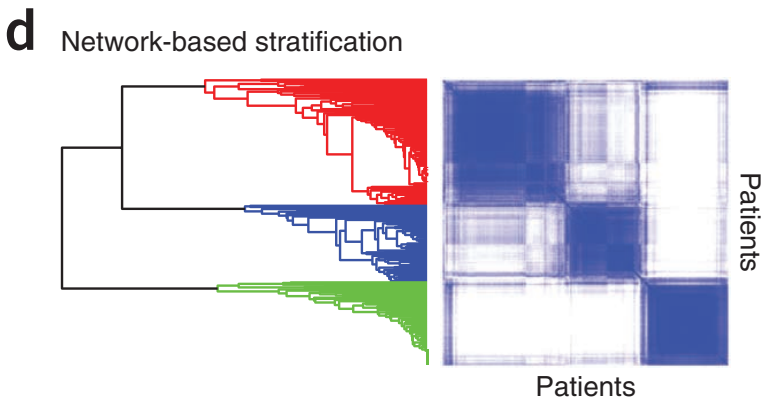
Matan Hofree[1], John P Shen[2], Hannah Carter[2], Andrew Gross[3] & Trey Ideker[1–3]

[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. [2]Department of Medicine, University of California, San Diego, La Jolla, California, USA. [3]Department 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





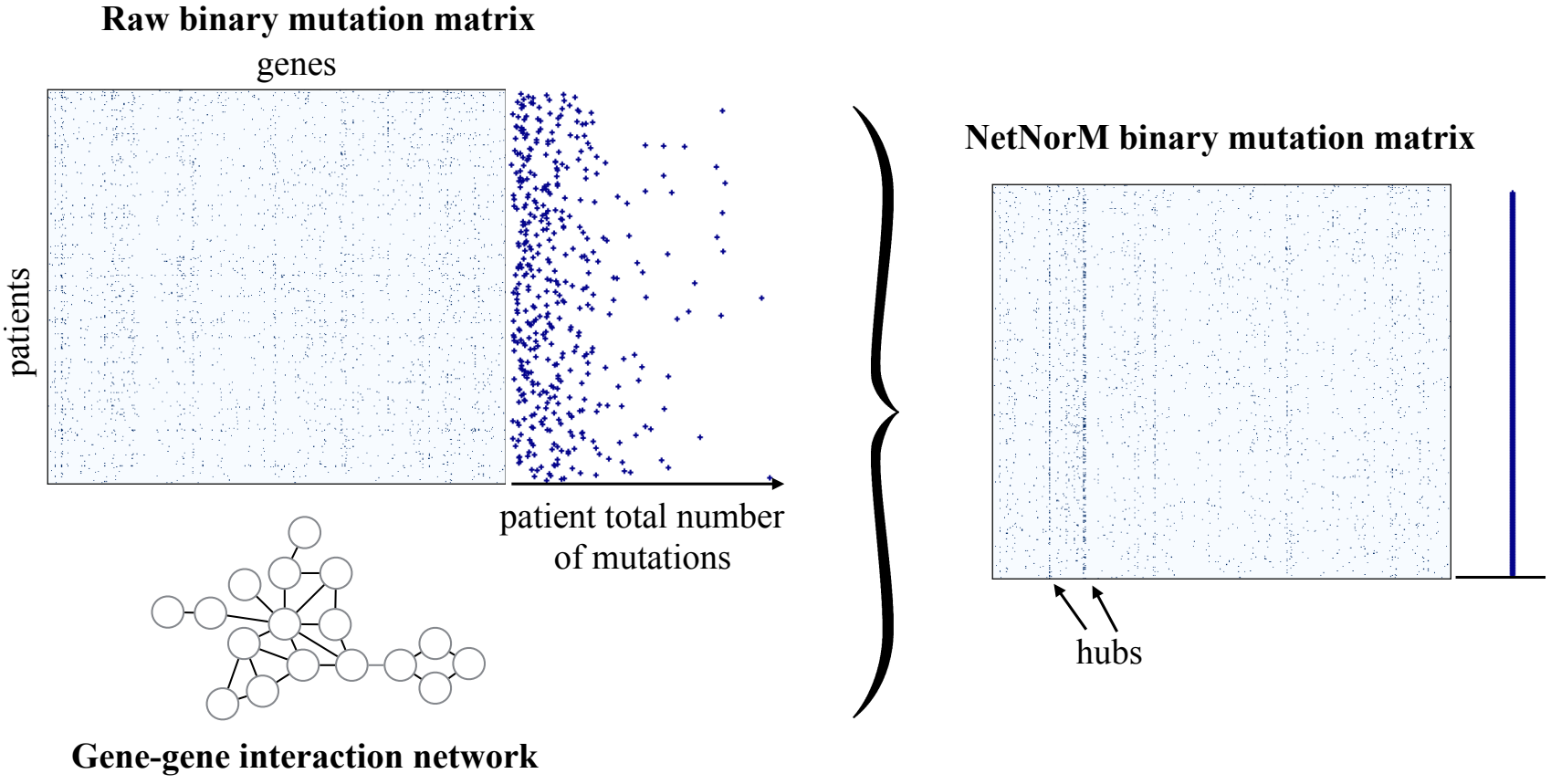


NetNorm Overview (Le Morvan et al., 2017)

Take

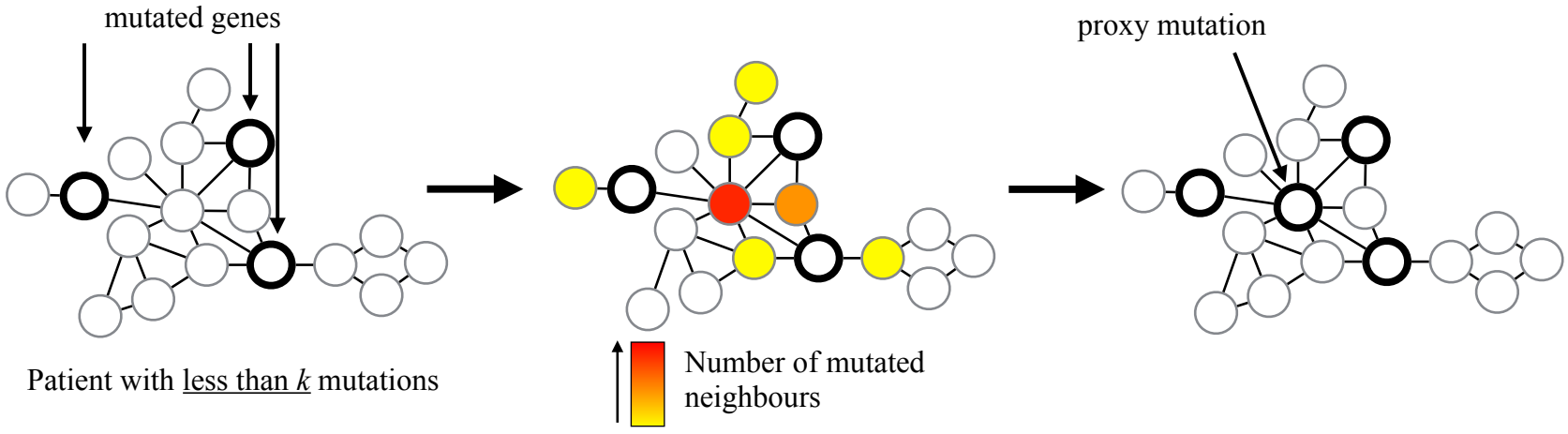
$$\mathcal{H} = \left\{ x \in \{0,1\}^p \, : \, \sum_{i=1}^{p} x_i = K \right\}$$

and use a gene network to transform x to $\phi(x) \in \mathcal{H}$ by adding/removing mutations

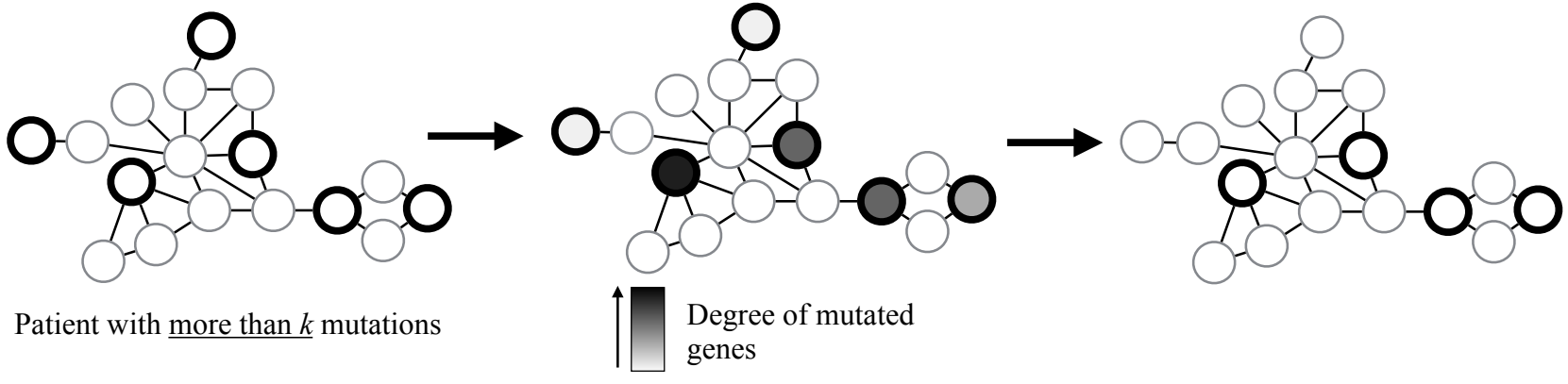


NetNorm detail (k=4)

1. Add mutations for patients with few (less than *K*) mutations

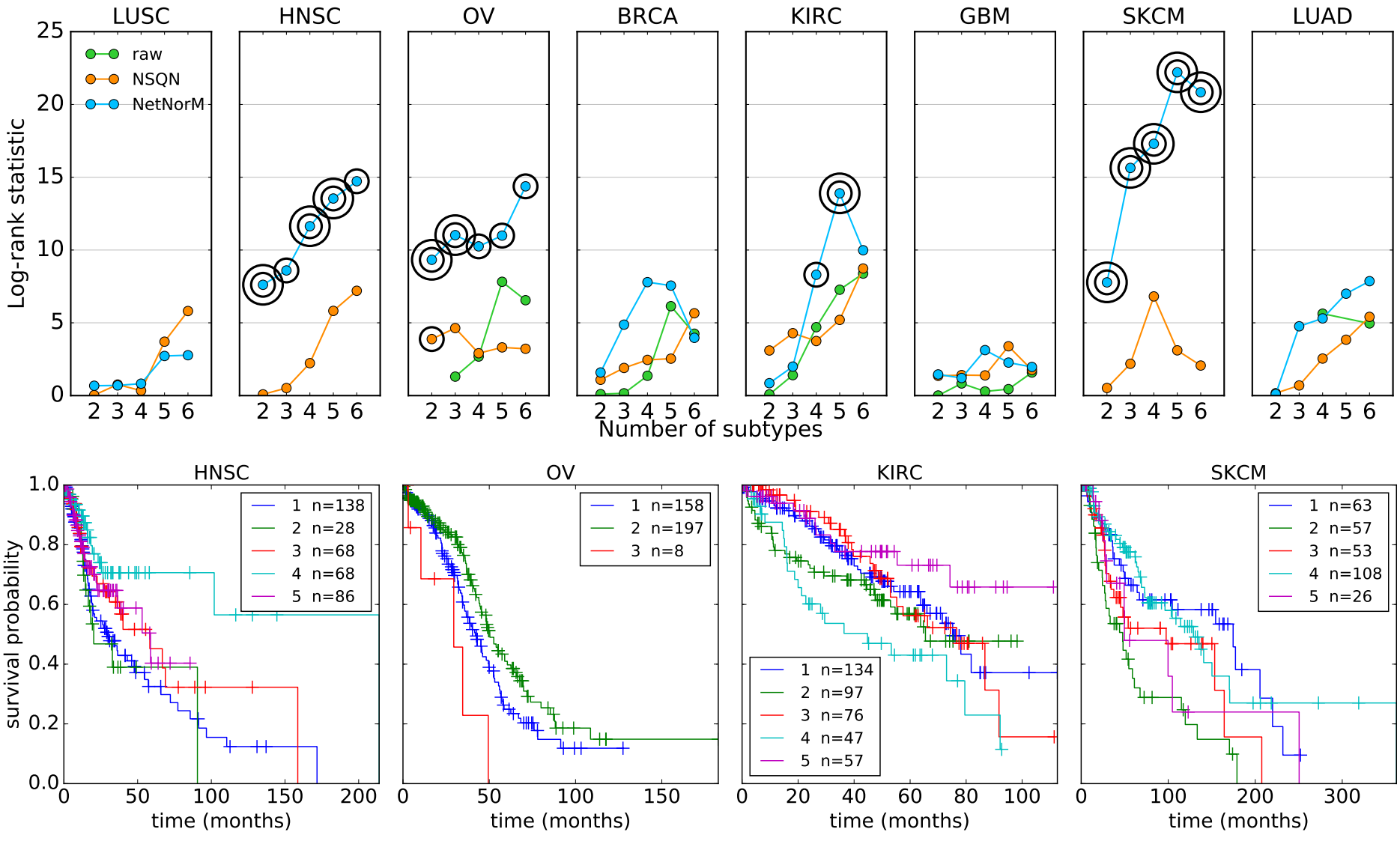


2. Remove mutations for patients for many (more than *K*) mutations

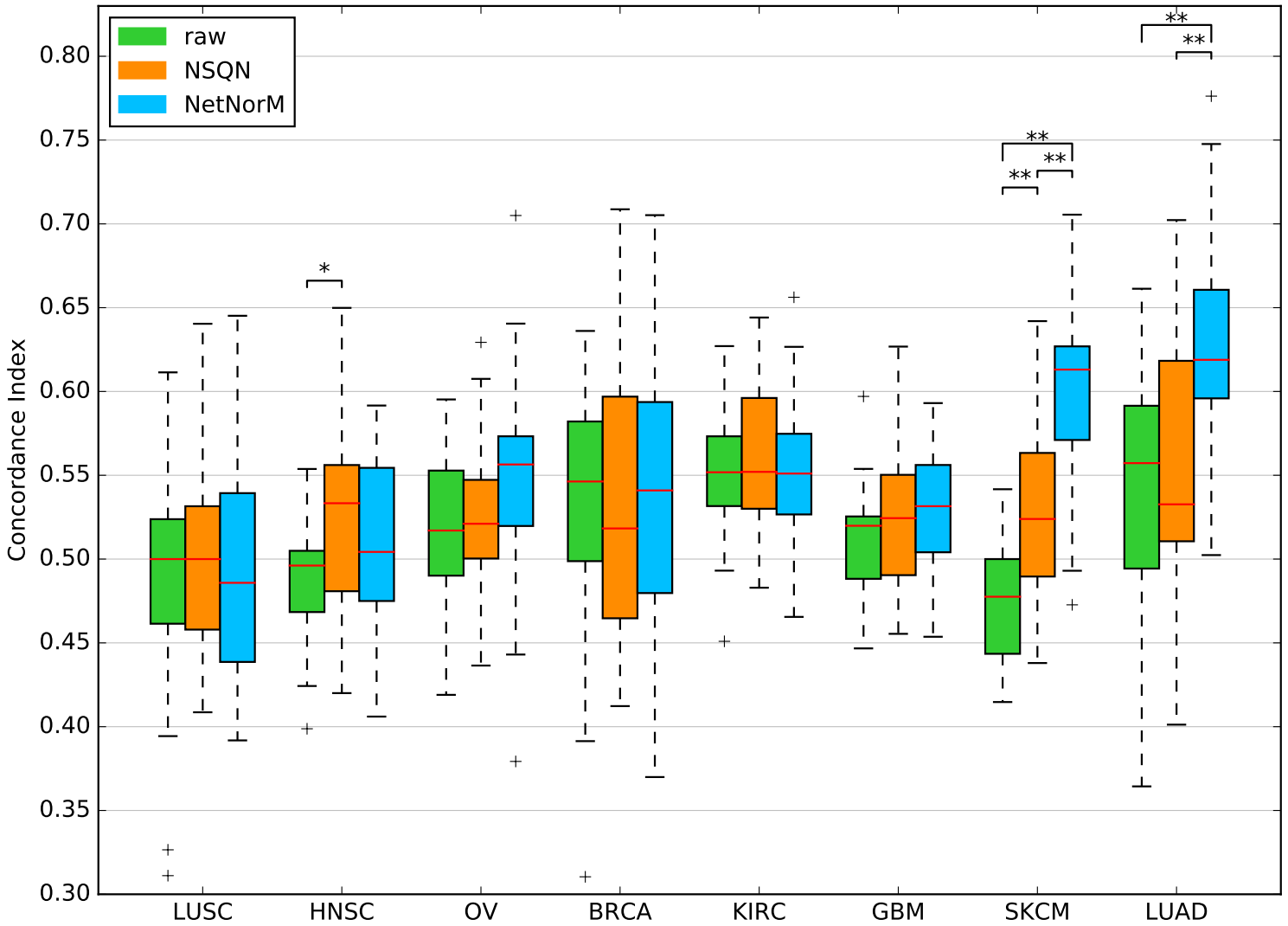


In practice, *K* is a free parameter optimized on the training set, typically a few 100's.

Results: unsupervised classification



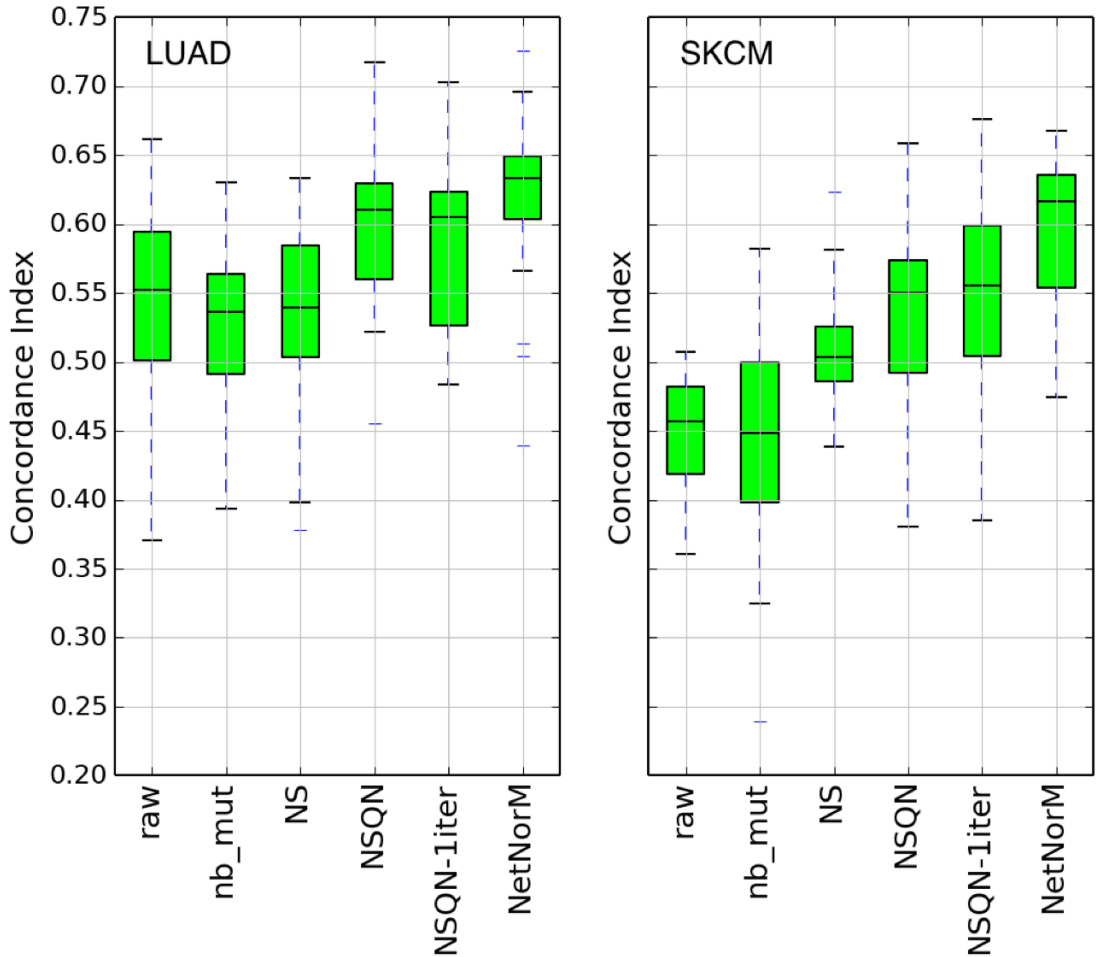
Results: survival prediction



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

The importance of Quantile Normalization (QN)

- Both NetNorm and NSQN transforms follow a 2-step a approach:
  1. Smooth the raw data onto the gene network (NS)
  2. Quantile normalize the smoothed profile (QN)
- QN matters!



Outline

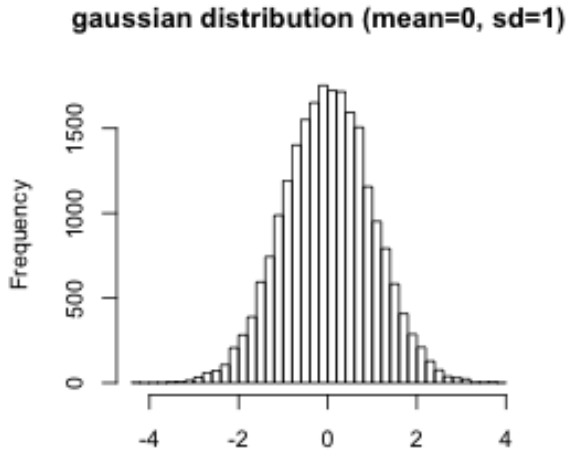
Standard QN

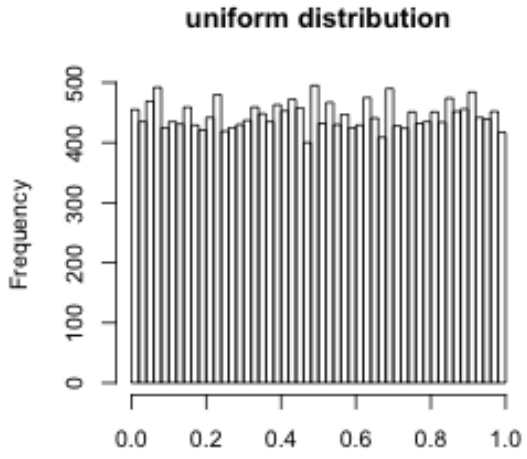
- Data: permutation $\sigma \in \mathbb{S}_n$ where $\sigma(i)$= rank of item/feature i
- Fix a target quantile $f \in \mathbb{R}^n$
- Define $\Phi_f : \mathbb{S}_n \to \mathbb{R}^n$ by

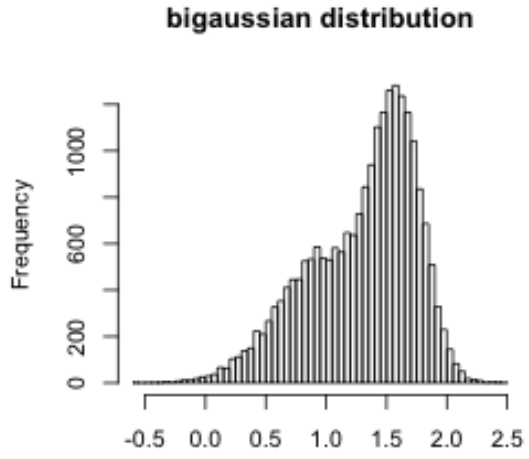
$$\forall \sigma \in \mathbb{S}_n\,, \quad [\Phi_f(\sigma)]_i = f_{\sigma(i)}$$

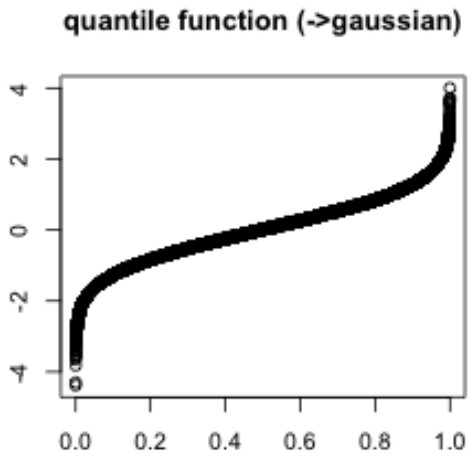
- "Keep the order, change the values"

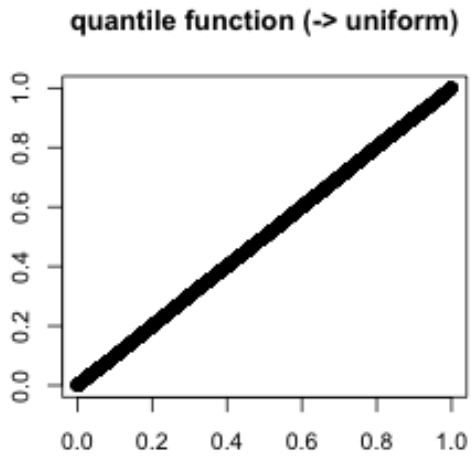
How to choose a "good" target distribution?

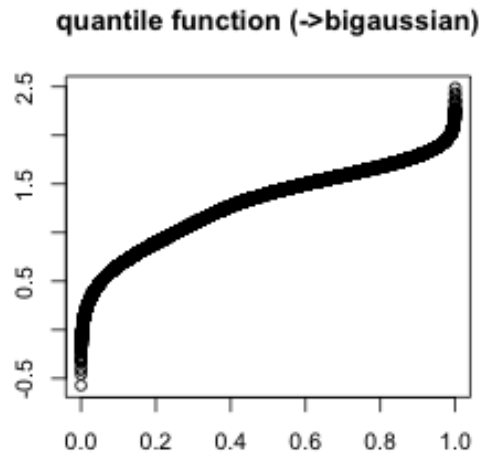












SUQUAN (Le Morvan and Vert, 2017)

Standard QN:

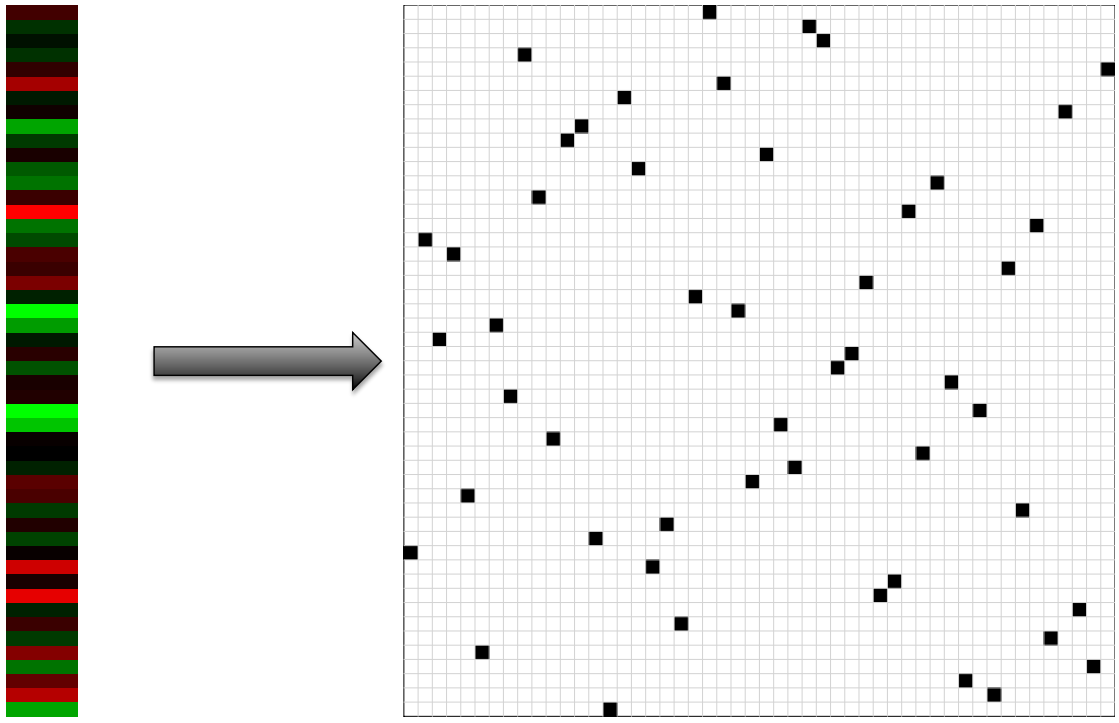
1. Fix f arbitrarily
2. QN all samples to get $\Phi_f(\sigma_1), \ldots, \Phi_f(\sigma_N)$
3. Learn a model on normalized data, e.g.:

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

Supervised QN (SUQUAN): jointly learn f and the model:

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

Computing $\Phi_f(\sigma)$



For $\sigma \in \mathbb{S}_n$ let the permutation representation (Serres, 1977):

$$[\Pi_\sigma]_{ij} = \begin{cases} 1 & \text{if } \sigma(j) = i\,, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\Phi_f(\sigma) = \Pi_\sigma^\top f$$

Linear SUQAN as rank-1 matrix regression

- Linear SUQUAN therefore solves

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

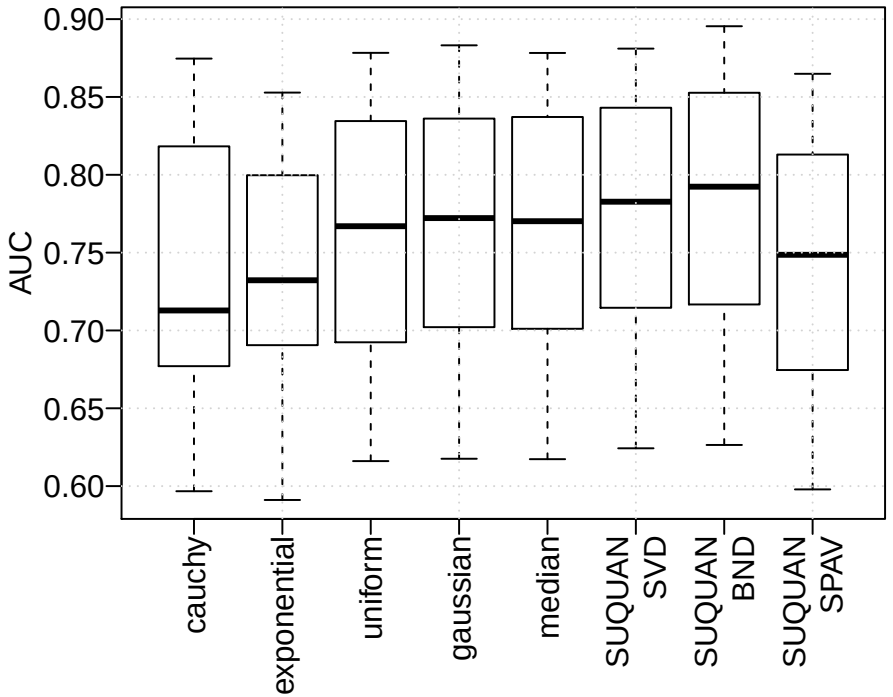
$$= \min_{w,b,f} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell \left(w^\top \Pi_{\sigma_i}^\top f + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}$$

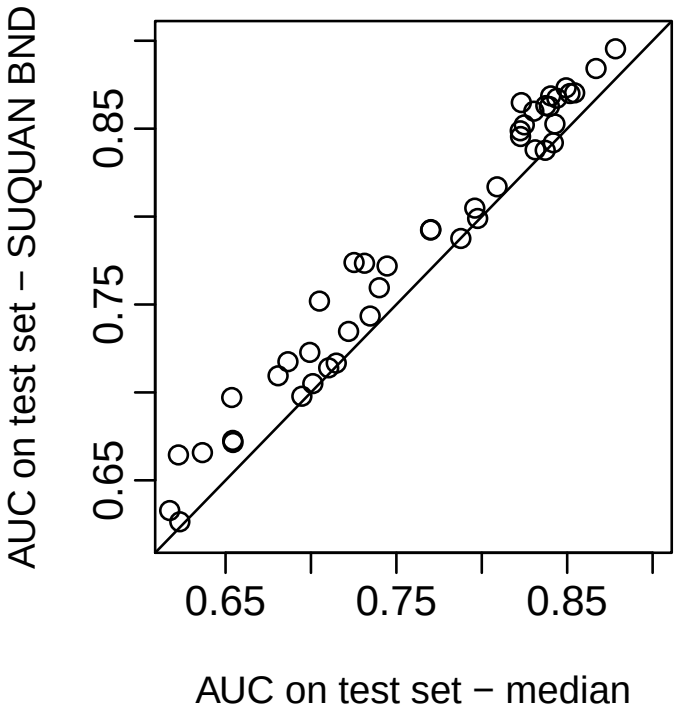
$$= \min_{w,b,f} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell \left(< \Pi_{\sigma_i}, f w^\top >_{\text{Frobenius}} + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}$$

- A particular linear model to estimate a rank-1 matrix $M = fw^\top$
- Each sample $\sigma \in \mathbb{S}_n$ is represented by the matrix $\Pi_\sigma \in \mathbb{R}^{n \times n}$
- Non-convex
- Alternative optimization of f and w is easy

Experiments: CIFAR-10

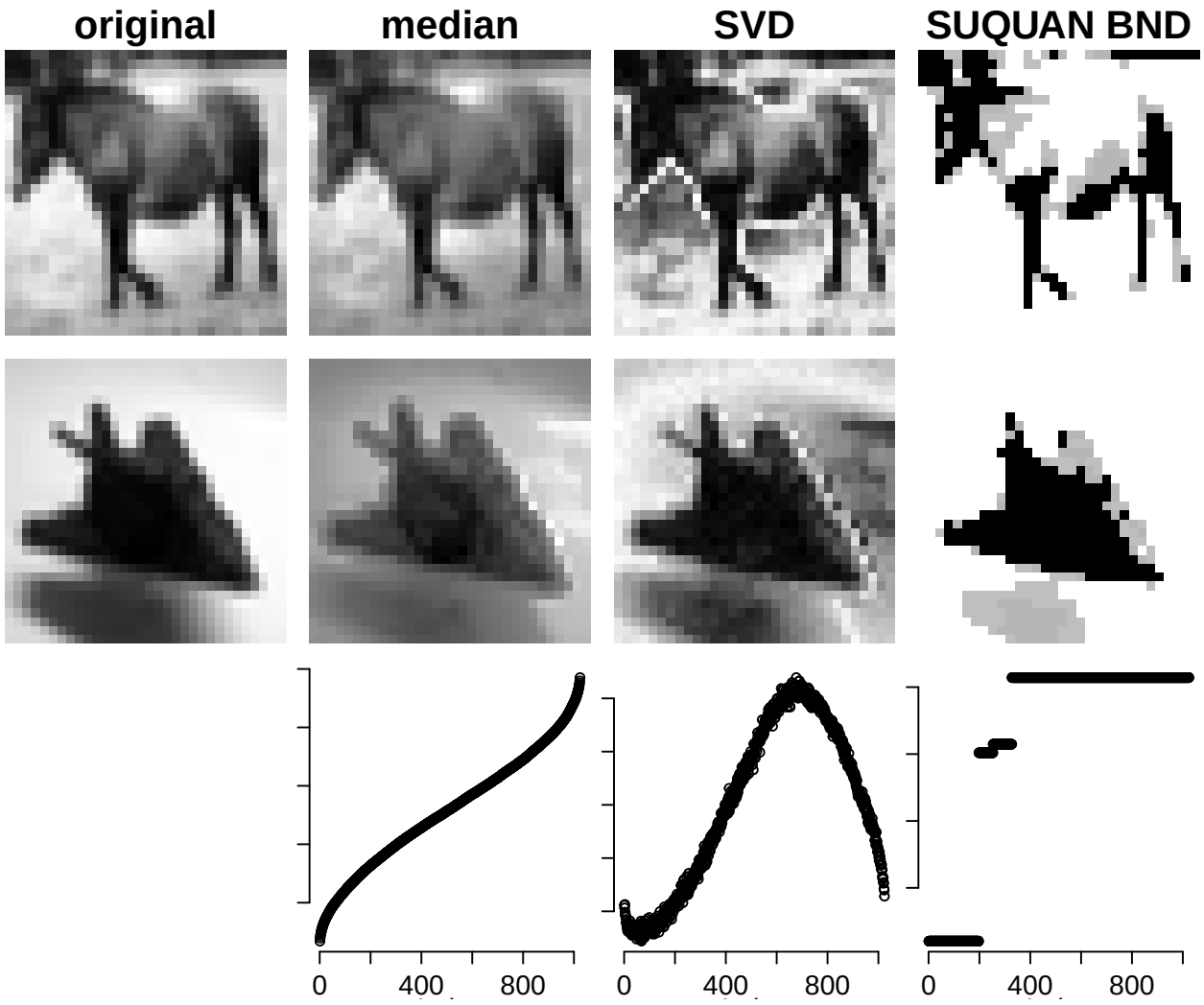
- Image classification into 10 classes (45 binary problems)
- $N = 5,000$ per class, $p = 1,024$ pixels





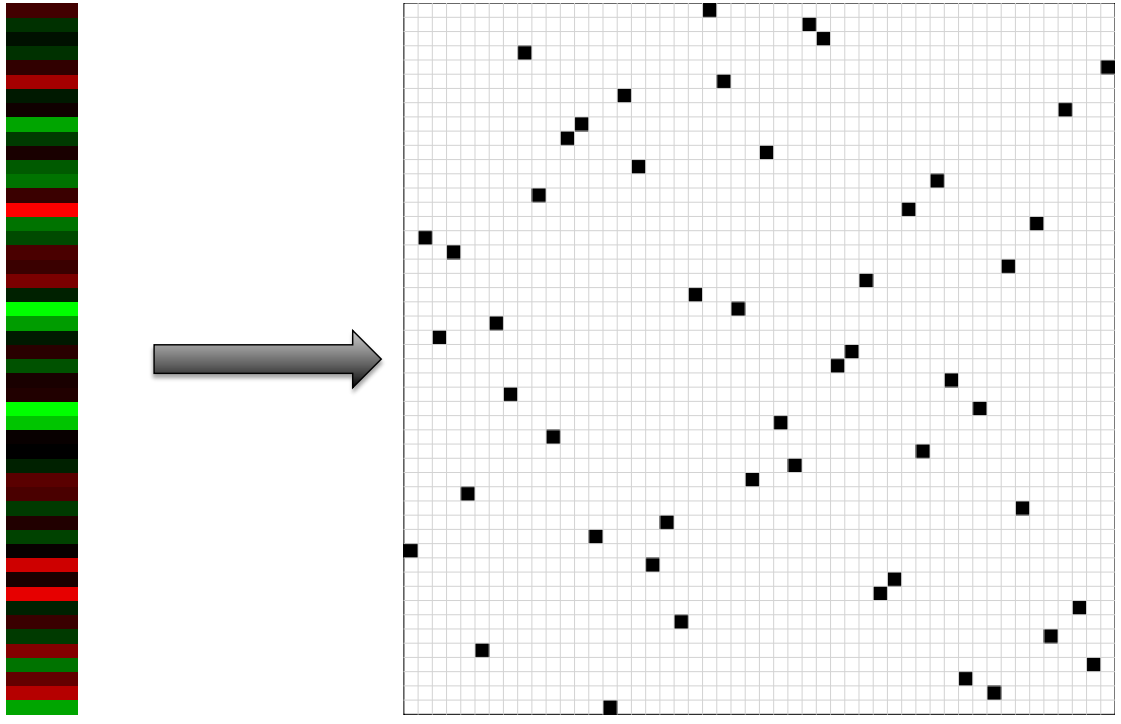
Experiments: CIFAR-10

- Example: horse vs. plane
- Different methods learn different quantile functions



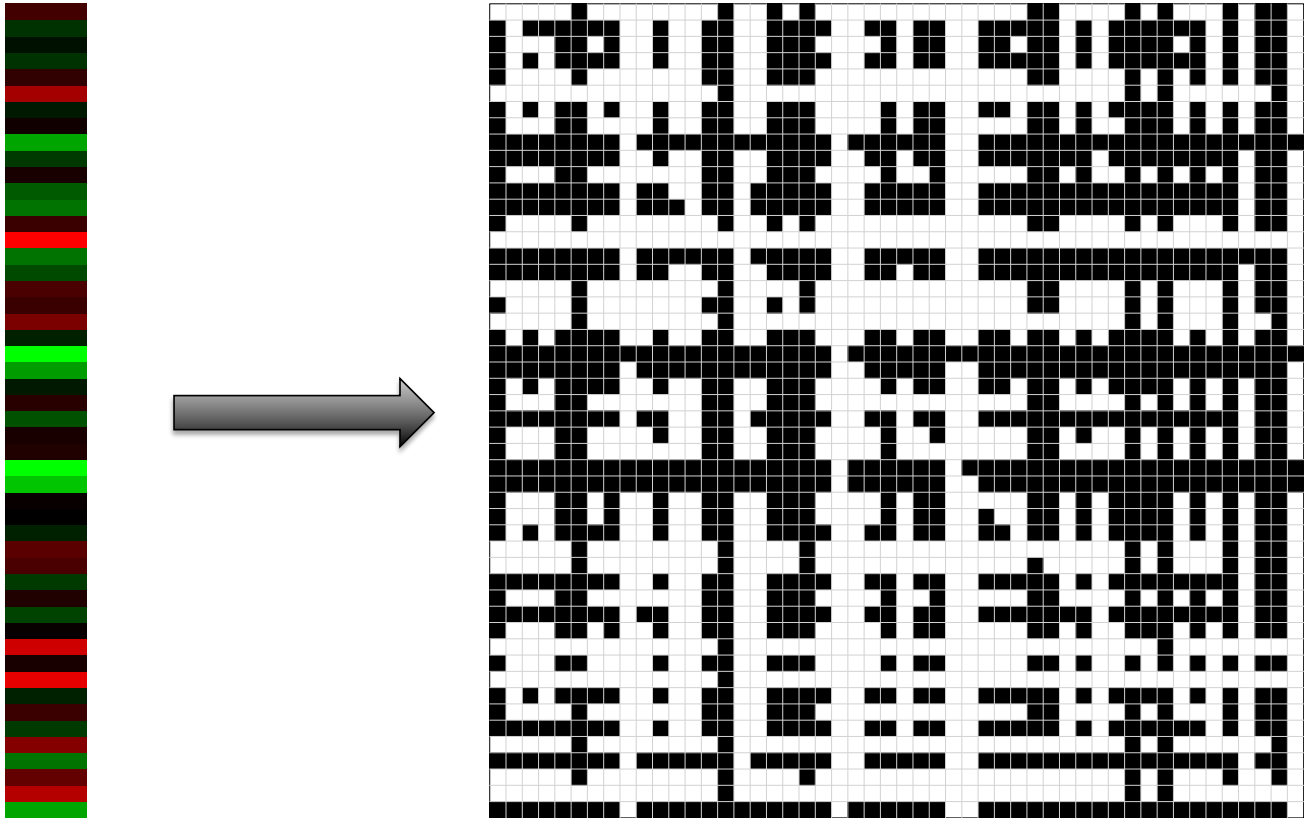
Outline

Limits of the QN embedding



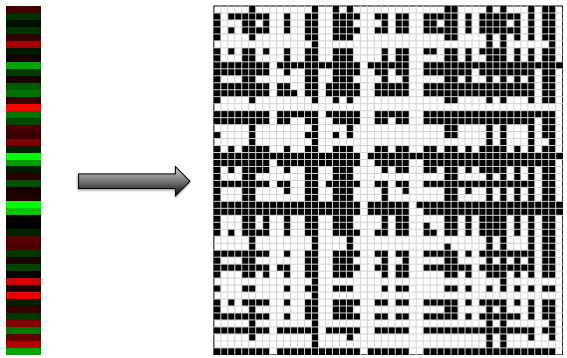
- Linear model on $\Phi(\sigma) = \Pi_\sigma \in \mathbb{R}^{n \times n}$
- Captures first-order information of the form "*i-th feature ranked at the j-th position*"
- What about higher-order information such as "*feature i larger than feature j*"?

Another representation



$$\Phi_{i,j}(\sigma) = \begin{cases} 1 & \text{if } \sigma(i) < \sigma(j)\,, \\ 0 & \text{otherwise.} \end{cases}$$

Geometry of the embedding



For any two permutations $\sigma, \sigma' \in \mathbb{S}_n$:

- Inner product

$$\Phi(\sigma)^\top \Phi(\sigma') = \sum_{1 \le i \ne j \le n} \mathbb{1}_{\sigma(i)<\sigma(j)} \mathbb{1}_{\sigma'(i)<\sigma'(j)} = n_c(\sigma, \sigma')$$

  n_c = number of concordant pairs

- Distance

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

  n_d = 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

$$\forall \lambda \geq 0 \quad K_M^\lambda(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')}$$

Theorem (Jiao and Vert, 2015, 2017)

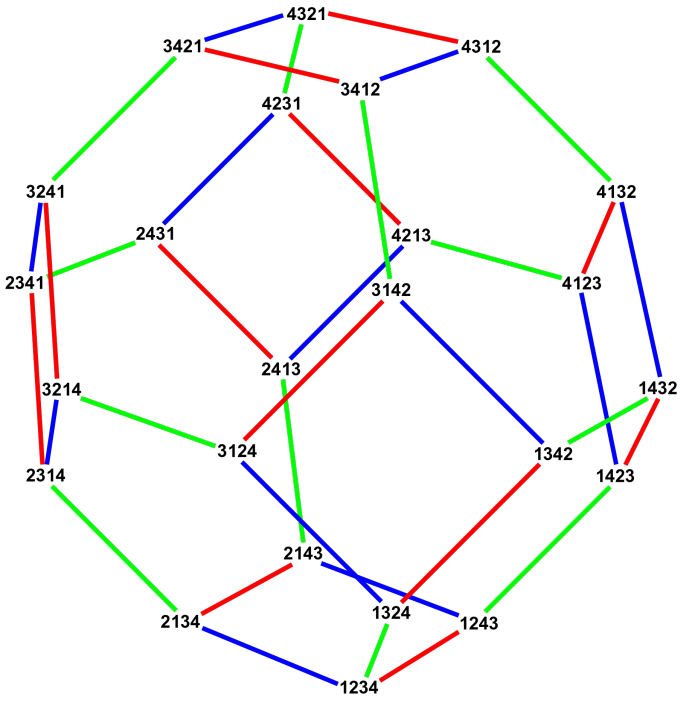
The Kendall and Mallows kernels are positive definite.

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

Related work



Cayley graph of $\mathbb{S}_4$

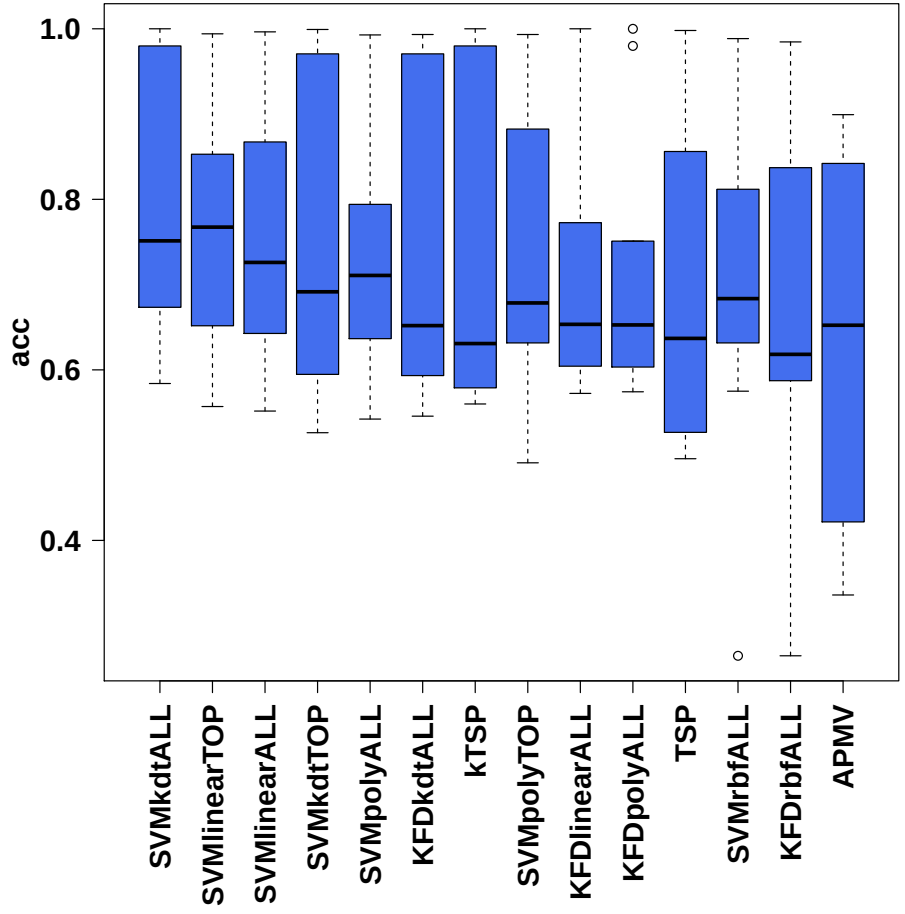
- Kondor and Barbarosa (2010) proposed the diffusion kernel on the Cayley graph of the symmetric group generated by adjacent transpositions.
- Computationally intensive ($O(n^{2n})$)
- Mallows kernel is written as

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

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

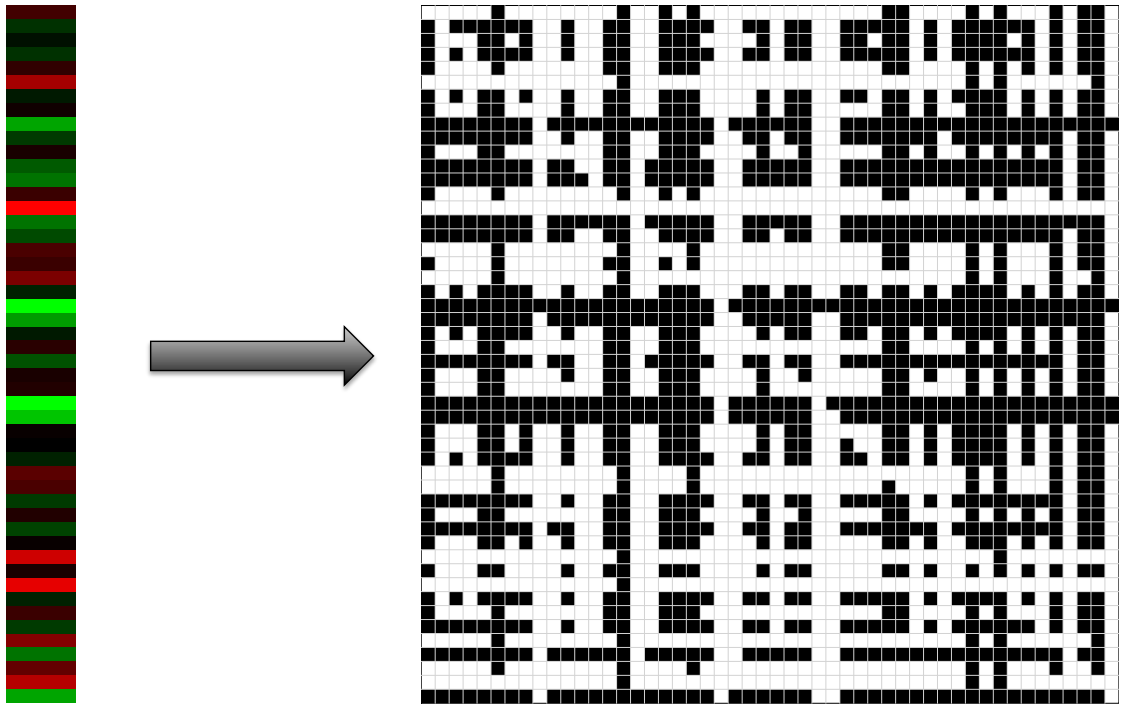
- It can be computed in $O(n \log n)$

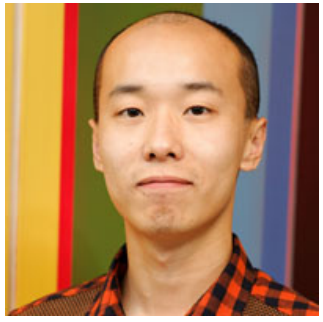
Applications



Average performance on 10 microarray classification problems (Jiao and Vert, 2017).

Extension: weighted Kendall kernel?





- Can we weight differently pairs based on their ranks?
- This would ensure a right-invariant kernel, i.e., the overall geometry does not change if we relabel the items

$$\forall \sigma_1, \sigma_2, \pi \in \mathbb{S}_n, \quad K(\sigma_1 \pi, \sigma_2 \pi) = K(\sigma_1, \sigma_2)$$

Related work

- Given a weight function $w : [1, n]^2 \to \mathbb{R}$, many weighted versions of the Kendall's τ have been proposed:

$$\sum_{1 \le i \ne j \le n} w(\sigma(i), \sigma(j)) \mathbb{1}_{\sigma(i) < \sigma(j)} \mathbb{1}_{\sigma'(i) < \sigma'(j)}$$

Shieh (1998)

$$\sum_{1 \le i \ne j \le n} w(\sigma(i), \sigma(j)) \frac{p_{\sigma(i)} - p_{\sigma'(i)}}{\sigma(i) - \sigma'(i)} \frac{p_{\sigma(j)} - p_{\sigma'(j)}}{\sigma(j) - \sigma'(j)} \mathbb{1}_{\sigma(i) < \sigma(j)} \mathbb{1}_{\sigma'(i) < \sigma'(j)}$$

Kumar and Vassilvitskii (2010)

$$\sum_{1 \le i \ne j \le n} w(i, j) \mathbb{1}_{\sigma(i) < \sigma(j)} \mathbb{1}_{\sigma'(i) < \sigma'(j)}$$

Vigna (2015)

- However, they are either not symmetric (1st and 2nd), or not right-invariant (3rd)

A right-invariant weighted Kendall kernel (Jiao and Vert, 2018)

Theorem

Let $W : \mathbb{N}^2 \times \mathbb{N}^2 \to \mathbb{R}$ *be a p.d. kernel on* $\mathbb{N}^2$*, then*

$$K_W(\sigma, \sigma') = \sum_{1 \leq i \neq j \leq n} W\left((\sigma(i), \sigma(j)), (\sigma'(i), \sigma'(j))\right) \mathbb{1}_{\sigma(i)<\sigma(j)} \mathbb{1}_{\sigma'(i)<\sigma'(j)}$$

is a right-invariant p.d. kernel on $\mathbb{S}_n$*.*

Corollary

For any matrix $U \in \mathbb{R}^{n \times n}$*,*

$$K_U(\sigma, \sigma') = \sum_{1 \leq i \neq j \leq n} U_{\sigma(i),\sigma(j)} U_{\sigma'(i),\sigma'(j)} \mathbb{1}_{\sigma(i)<\sigma(j)} \mathbb{1}_{\sigma'(i)<\sigma'(j)} ,$$

is a right-invariant p.d. kernel on $\mathbb{S}_n$*.*

A right-invariant weighted Kendall kernel (Jiao and Vert, 2018)

Theorem

Let $W : \mathbb{N}^2 \times \mathbb{N}^2 \to \mathbb{R}$ *be a p.d. kernel on* $\mathbb{N}^2$*, then*

$$K_W(\sigma, \sigma') = \sum_{1 \leq i \neq j \leq n} W\left((\sigma(i), \sigma(j)), (\sigma'(i), \sigma'(j))\right) \mathbb{1}_{\sigma(i)<\sigma(j)} \mathbb{1}_{\sigma'(i)<\sigma'(j)}$$

is a right-invariant p.d. kernel on $\mathbb{S}_n$*.*

Corollary

For any matrix $U \in \mathbb{R}^{n \times n}$*,*

$$K_U(\sigma, \sigma') = \sum_{1 \leq i \neq j \leq n} U_{\sigma(i),\sigma(j)} U_{\sigma'(i),\sigma'(j)} \mathbb{1}_{\sigma(i)<\sigma(j)} \mathbb{1}_{\sigma'(i)<\sigma'(j)} \,,$$

is a right-invariant p.d. kernel on $\mathbb{S}_n$*.*

Examples

$U_{a,b}$ corresponds to the weight of (items ranked at) positions a and b in a permutation. Interesting choices include:

- *Top-k.* For some $k \in [1, n]$,

$$U_{a,b} = \begin{cases} 1 & \text{if } a \leq k \text{ and } b \leq k\,, \\ 0 & \text{otherwise.} \end{cases}$$

- *Additive.* For some $u \in \mathbb{R}^n$, take

$$U_{ij} = u_i + u_j$$

- *Multiplicative.* For some $u \in \mathbb{R}^n$, take

$$U_{ij} = u_i u_j$$

Theorem (Kernel trick)

The weighted Kendall kernel can be computed in $O(n \ln(n))$ for the top-k, additive or multiplicative weights.

Learning the weights (1/2)

- K_U can be written as

$$K_U(\sigma, \sigma') = \Phi_U(\sigma)^\top \Phi_U(\sigma')$$

with

$$\Phi_U(\sigma) = \left(U_{\sigma(i),\sigma(j)} \mathbb{1}_{\sigma(i)<\sigma(j)}\right)_{1 \leq i \neq j \leq n}$$

- Interesting fact: For any upper triangular matrix $U \in \mathbb{R}^{n \times n}$,

$$\Phi_U(\sigma) = \Pi_\sigma^\top U \Pi_\sigma \quad \text{with } (\Pi_\sigma)_{ij} = \mathbb{1}_{i=\sigma(j)}$$

- Hence a linear model on Φ_U can be rewritten as

$$\begin{aligned} f_{\beta,U}(\sigma) &= \langle \beta, \Phi_U(\sigma) \rangle_{\text{Frobenius}(n \times n)} \\ &= \left\langle \beta, \Pi_\sigma^\top U \Pi_\sigma \right\rangle_{\text{Frobenius}(n \times n)} \\ &= \left\langle \Pi_\sigma \otimes \Pi_\sigma, \text{vec}(U) \otimes (\text{vec}(\beta))^\top \right\rangle_{\text{Frobenius}(n^2 \times n^2)} \end{aligned}$$

Learning the weights (2/2)

$$f_{\beta,U}(\sigma) = \left\langle \Pi_\sigma \otimes \Pi_\sigma, \mathrm{vec}(U) \otimes (\mathrm{vec}(\beta))^\top \right\rangle_{\text{Frobenius}(n^2 \times n^2)}$$

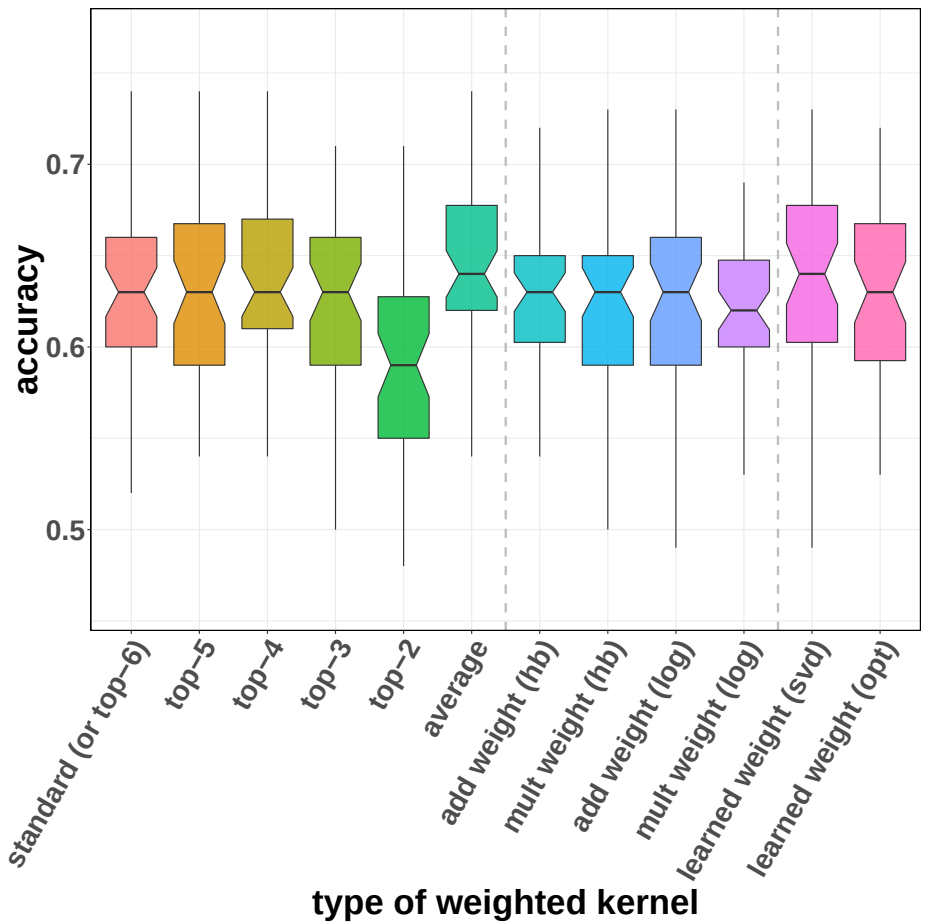
- This is symmetric in U and β
- Instead of fixing the weights U and optimizing β, we can jointly optimize β and U to learn the weights U
- Note that $\Pi_\sigma^\top = (\Pi_\sigma)^{-1} = \Pi_{\sigma^{-1}}$, hence

$$f_{\beta,U}(\sigma) = f_{U,\beta}(\sigma^{-1})$$

- We propose to alternate optimization in U and β
  - For U fixed, optimize β with $K_U(\sigma_1, \sigma_2)$
  - For β fixed, optimize U with $K_\beta(\sigma_1^{-1}, \sigma_2^{-1})$

Experiments

- Eurobarometer data (Christensen, 2010)
- >12k individuals rank 6 sources of information
- Binary classification problem: predict age from ranking (>40y vs <40y)



Weights learned

Towards higher-order representations

$$f_{\beta,U}(\sigma) = \left\langle \Pi_\sigma \otimes \Pi_\sigma, \mathrm{vec}(U) \otimes (\mathrm{vec}(\beta))^\top \right\rangle_{\text{Frobenius}(n^2 \times n^2)}$$

- A particular rank-1 linear model for the embedding

$$\Sigma_\sigma = \Pi_\sigma \otimes \Pi_\sigma \in (\{0,1\})^{n^2 \times n^2}$$

- Σ is the direct sum of the second-order and first-order permutation representations:

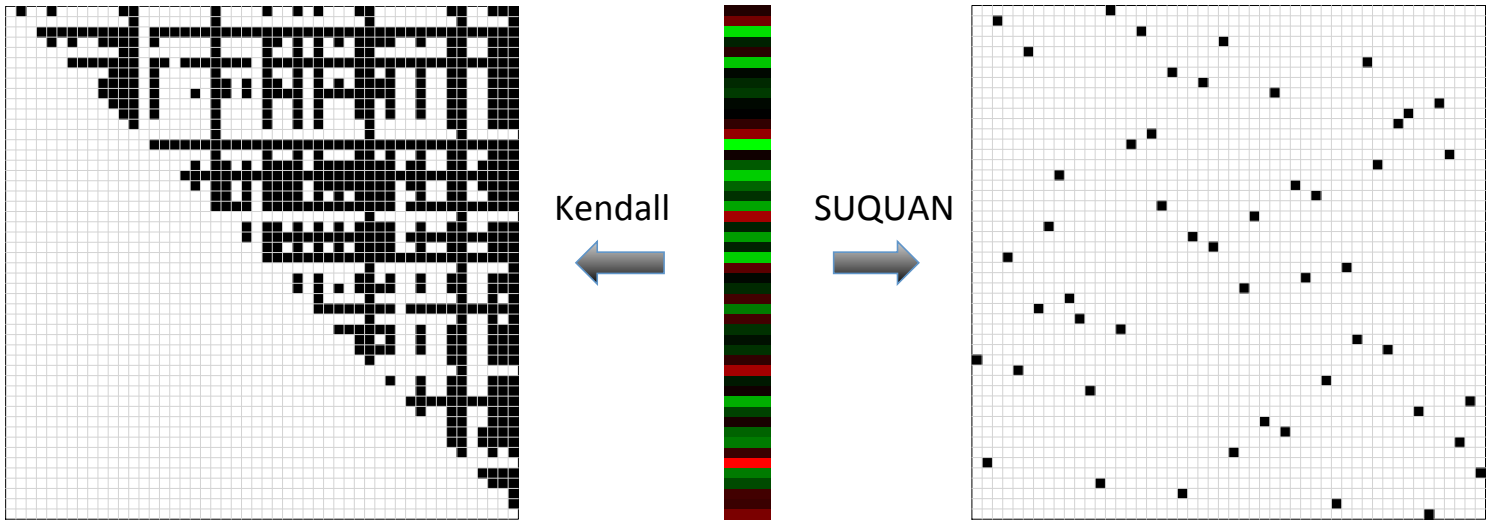
$$\Sigma \cong \tau_{(n-2,1,1)} \oplus \tau_{(n-1,1)}$$

- This generalizes SUQUAN which considers the first-order representation Π_σ only:

$$h_{\beta,w}(\sigma) = \left\langle \Pi_\sigma, w \otimes \beta^\top \right\rangle_{\text{Frobenius}(n \times n)}$$

- Generalization possible to higher-order information by using higher-order linear representations of the symmetric group, which are the good basis for right-invariant kernels (Bochner theorem)...

Conclusion



- Lots of complex data in genomics; feature engineering still relevant
- Machine learning beyond vectors, strings and graphs
- Different embeddings of the symmetric group
- Respect the group structure (right-invariance) through group representations
- Compatible with NN architectures
- Scalability? Approximate embeddings?

Thanks

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

References

R. E. Barlow, D. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions; the theory and application of isotonic regression*. Wiley, New-York, 1972.

T. Christensen. Eurobarometer 55.2: Science and technology, agriculture, the euro, and internet access, may-june 2001. https://doi.org/10.3886/ICPSR03341.v3, June 2010. ICPSR03341-v3. Cologne, Germany: GESIS/Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 2010-06-30.

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`.

Y. Jiao and J.-P. Vert. The Kendall and Mallows kernels for permutations. In *Proceedings of The 32nd International Conference on Machine Learning*, volume 37 of *JMLR:W&CP*, pages 1935–1944, 2015. URL `http://jmlr.org/proceedings/papers/v37/jiao15.html`.

Y. Jiao and J.-P. Vert. The Kendall and Mallows kernels for permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. doi: 10.1109/TPAMI.2017.2719680. URL `http://dx.doi.org/10.1109/TPAMI.2017.2719680`.

Y. Jiao and J.-P. Vert. The weighted kendall and high-order kernels for permutations. Technical Report 1802.08526, arXiv, 2018.

W. R. Knight. A computer method for calculating Kendall's tau with ungrouped data. *J. Am. Stat. Assoc.*, 61(314):436–439, 1966. URL `http://www.jstor.org/stable/2282833`.

References (cont.)

R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web (WWW-10)*, pages 571–580. ACM, 2010. doi: 10.1145/1772690.1772749.

M. Le Morvan and J.-P. Vert. Supervised quantile normalisation. Technical Report 1706.00244, arXiv, 2017.

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. *PLoS Comp. Bio.*, 13(6):e1005573, 2017. URL http://hal.archives-ouvertes.fr/hal-01341856.

J.-P. Serres. *Linear Representations of Finite Groups*. Graduate Texts in Mathematics. Springer-Verlag New York, 1977. doi: 10.1007/978-1-4684-9458-7. URL http://dx.doi.org/10.1007/978-1-4684-9458-7.

G. S. Shieh. A weighted Kendall's tau statistic. *Statistics & Probability Letters*, 39(1):17–24, 1998. doi: 10.1016/s0167-7152(98)00006-6. URL http://dx.doi.org/10.1016/S0167-7152(98)00006-6.

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.

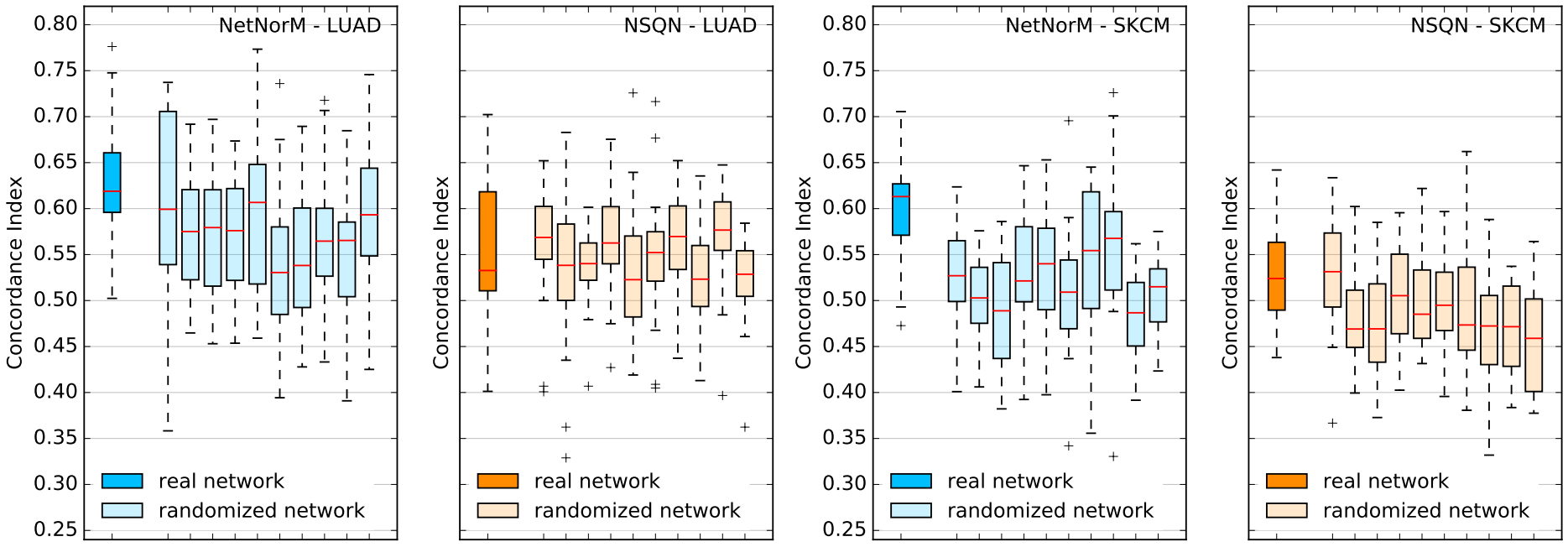
O. Sysoev and O. Burdakov. A smoothed monotonic regression via l2 regularization. Technical Report LiTH-MAT-R–2016/01–SE, Department of mathematics, Linköping University, 2016. URL http://liu.diva-portal.org/smash/get/diva2:905380/FULLTEXT01.pdf.

References (cont.)

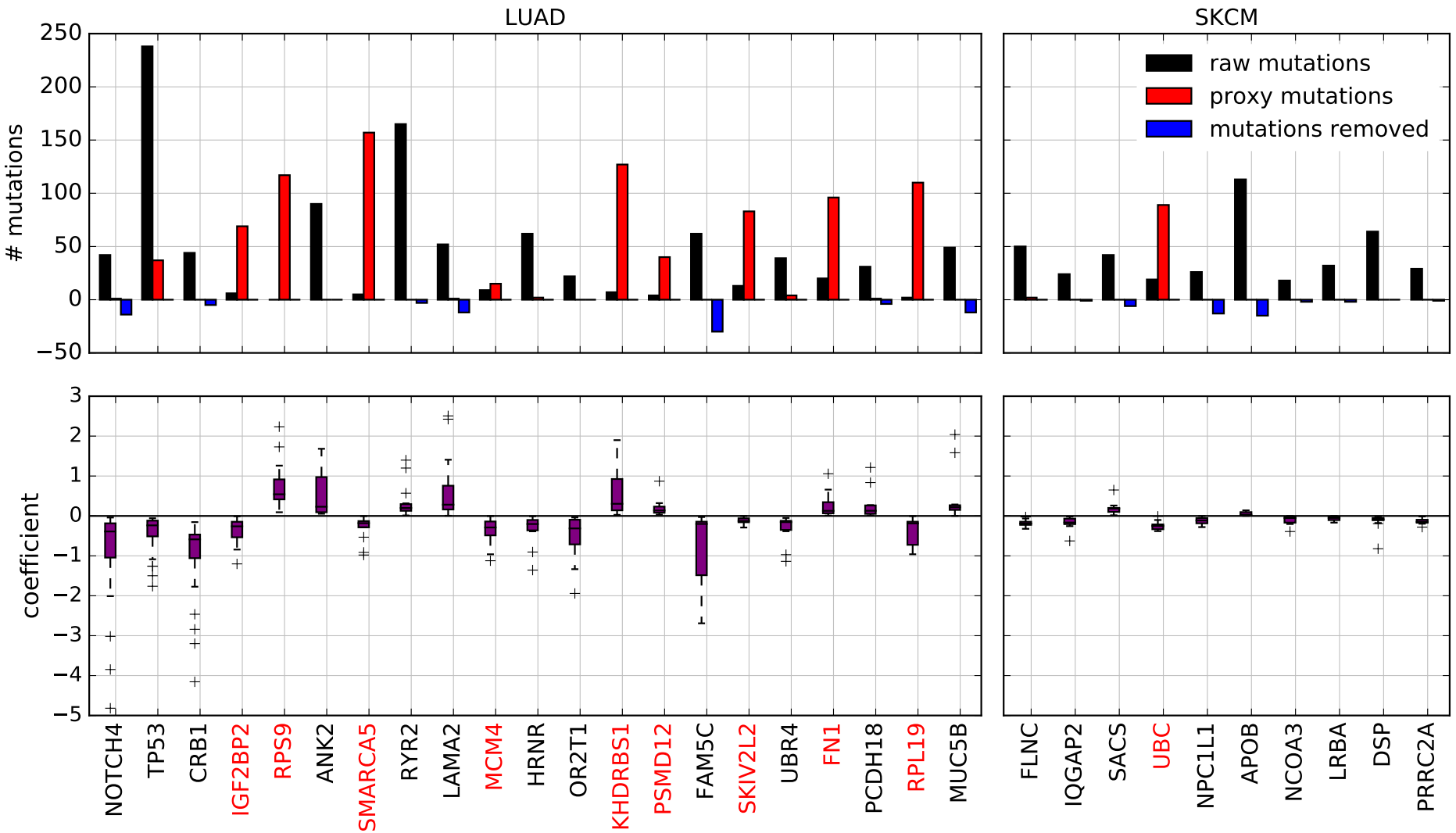
S. Vigna. A weighted correlation index for rankings with ties. In *Proceedings of the 24th International Conference on World Wide Web (WWW-15)*, pages 1166–1176. ACM, 2015. doi: 10.1145/2736277.2741088.

NetNorM and NSQN benefit from biological information in the gene network

Comparison with 10 randomly permuted networks:

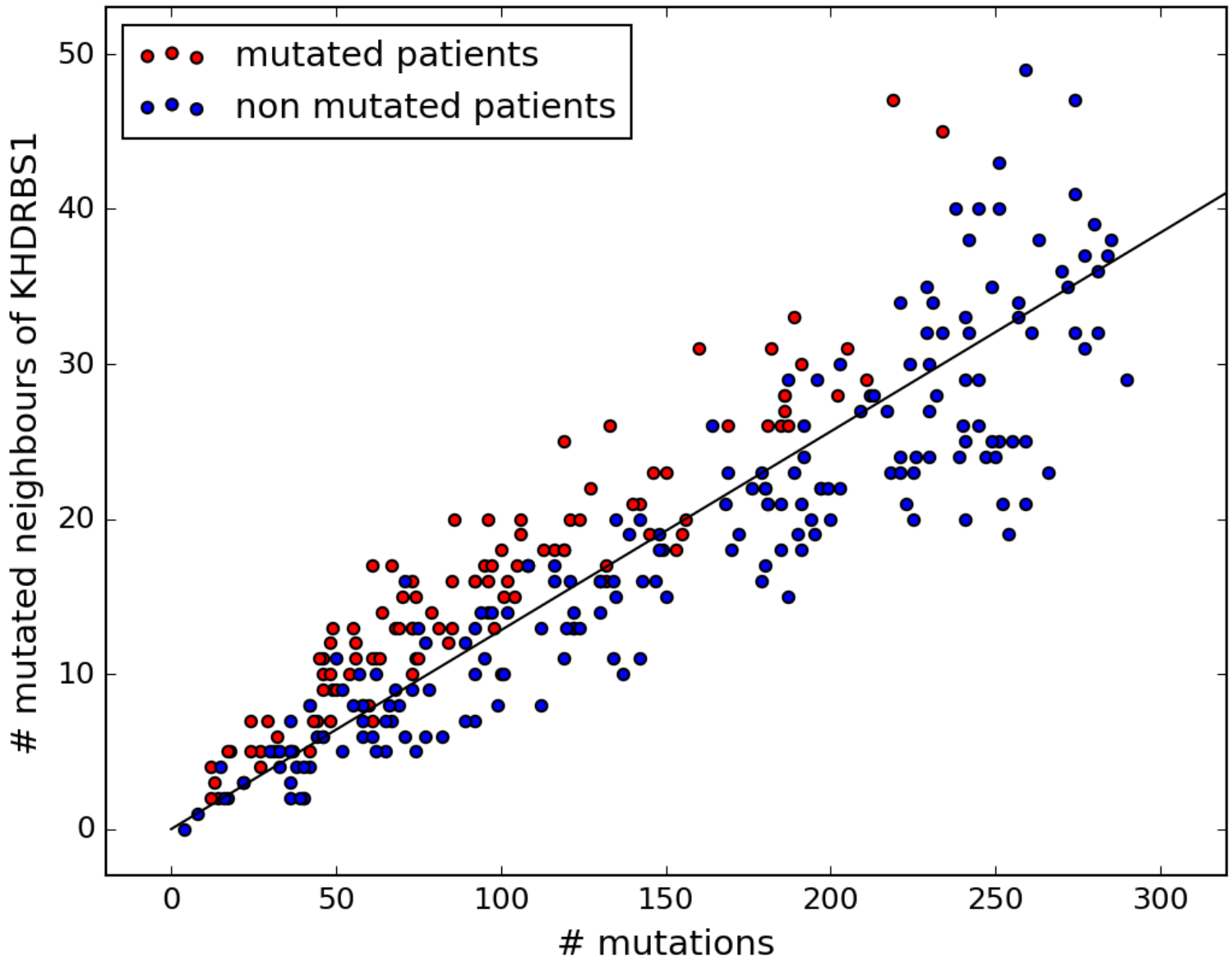


Selected genes represent "true" or "proxy" mutations

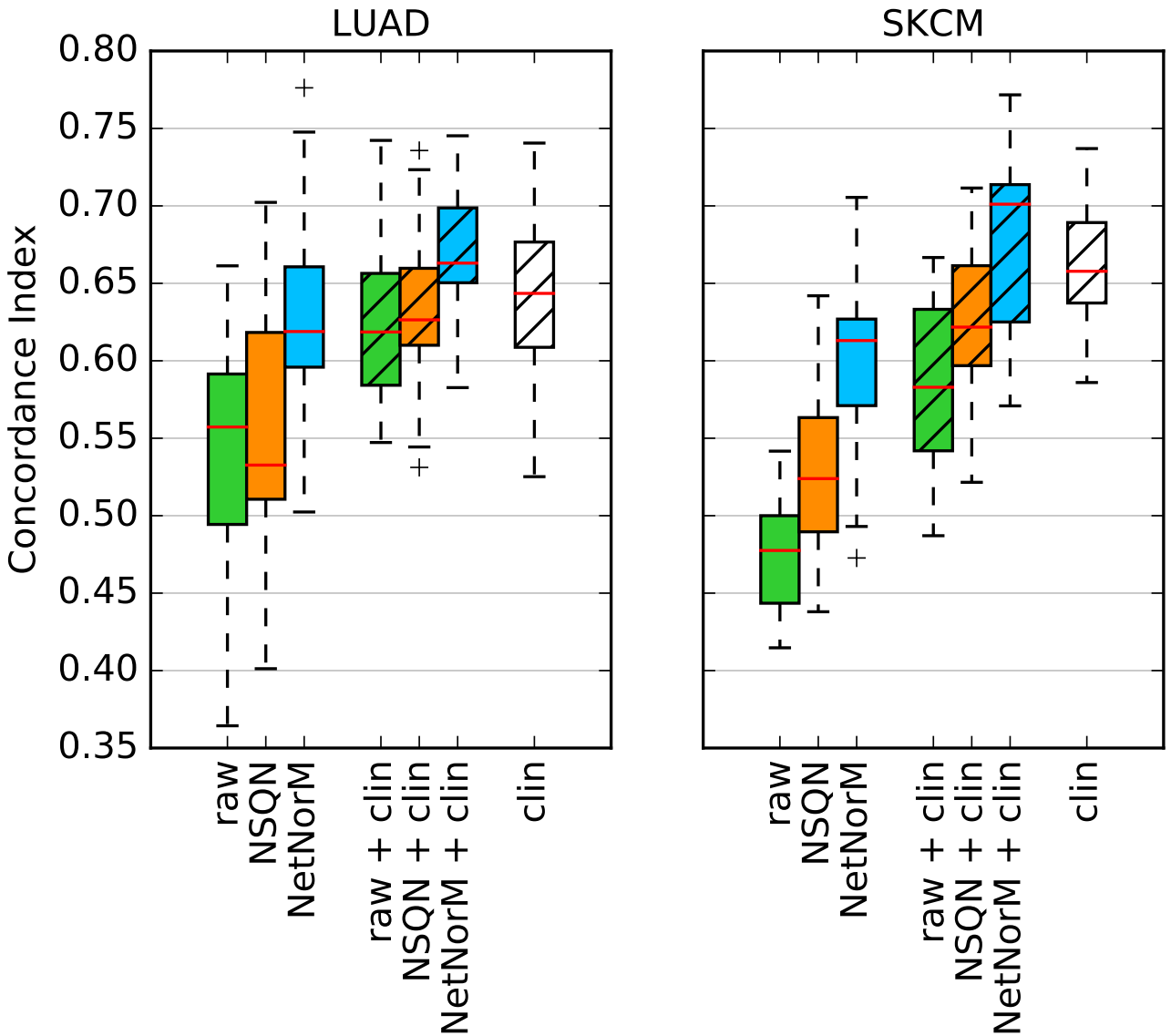


Genes selected in at least 50% of the cross-validated sparse SVM model

Proxy mutations encode both total number of mutations and local mutational burden

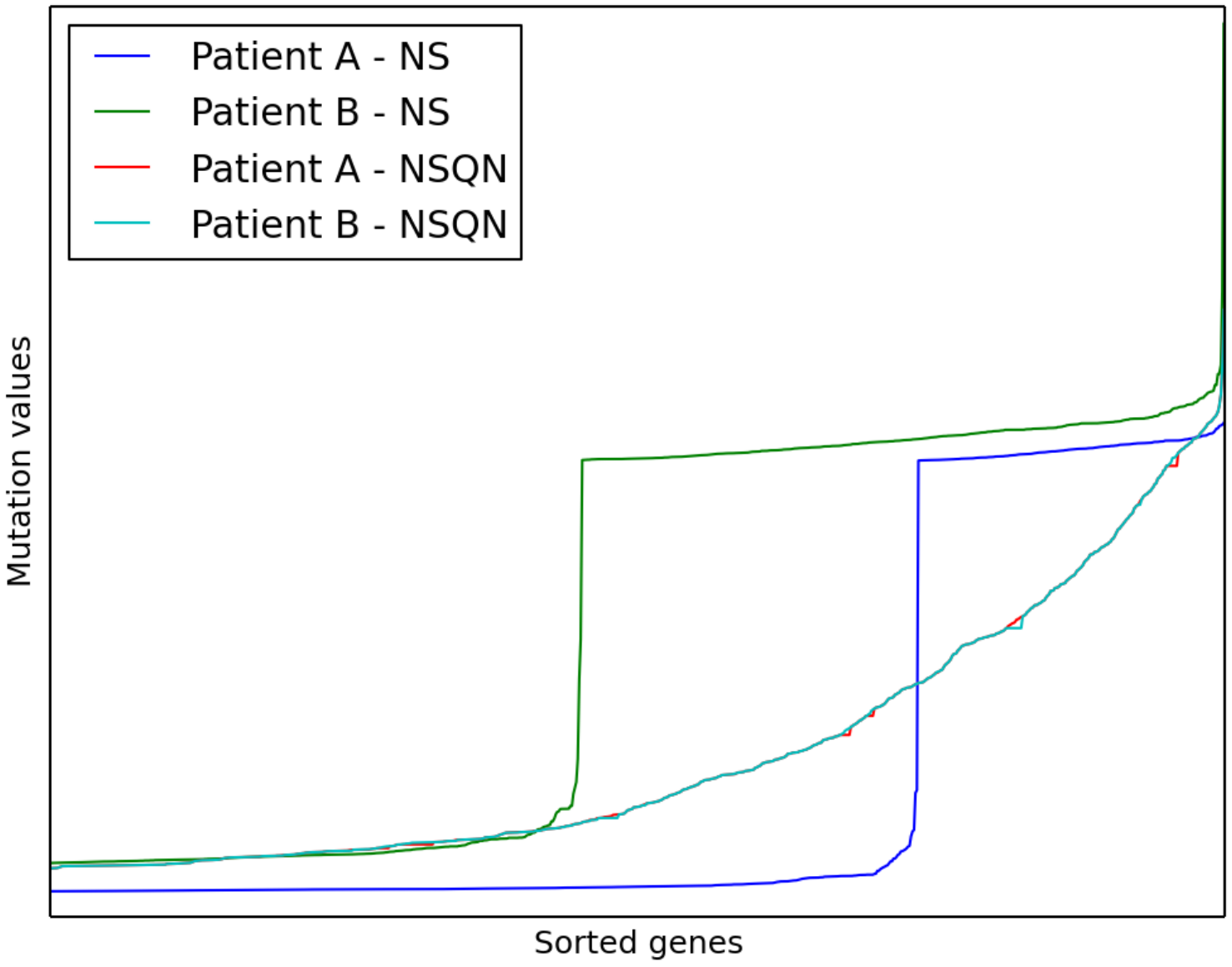


Adding good old clinical factors



Combination by averaging predictions

QN after network smoothing



Constraints on f

- Ridge

$$\mathcal{F}_0 = \left\{ f \in \mathbb{R}^p \,:\, \frac{1}{p}\sum_{i=1}^{p} f_i^2 \leq 1 \right\} .$$

- Non-decreasing

$$\mathcal{F}_{\text{BND}} = \mathcal{F}_0 \cap \mathcal{I}_0 \,, \quad \text{where} \quad \mathcal{I}_0 = \{ f \in \mathbb{R}^p \,:\, f_1 \leq f_2 \leq \ldots \leq f_p \}$$

- Non-decreasing and smooth

$$\mathcal{F}_{\text{SPAV}} = \left\{ f \in \mathcal{I}_0 \,:\, \sum_{j=1}^{p-1} (f_{j+1} - f_j)^2 \leq 1 \right\} .$$

SUQUAN-BND and SUQUAN-PAVA

Algorithm 2: SUQUAN-BND and SUQUAN-SPAV

Input: $(x_1, y_1), \ldots, (x_n, y_n), f_{init} \in \mathcal{I}_0, \lambda \in \mathbb{R}$

Output: $f \in \mathcal{I}_0$ target quantile

1: **for** $i = 1$ to n **do**

2: $\quad rank_i, order_i \leftarrow \text{sort}(x_i)$

3: **end for**

4: $w, b \leftarrow \underset{w,b}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_i \left(w^\top f_{init}[rank_i] + b\right) + \lambda ||w||^2$

(standard linear model optimisation)

5: $f \leftarrow \underset{f \in \mathcal{F}_{BND}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_i \left(f^\top w[order_i] + b\right)$

(isotonic optimisation problem using PAVA as prox)

OR

$f \leftarrow \underset{f \in \mathcal{F}_{SPAV}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_i \left(f^\top w[order_i] + b\right)$

(smoothed isotonic optimisation problem using SPAV as prox)

- Alternate optimization in *w* and *f*, monotonicity constraint on *f*
- Accelerated proximal gradient optimization for *f*, using the Pool Adjacent Violators Algorithm (PAVA, Barlow et al. (1972)) or the Smoothed Pool Adjacent Violators algorithm (SPAV, Sysoev and Burdakov (2016)) as proximal operator.

A variant: SUQUAN-SVD

Algorithm 1: SUQUAN-SVD

Input:

$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, 1\}$

Output: $f \in \mathcal{F}_0$ target quantile

1: $M_{LDA} \leftarrow 0 \in \mathbb{R}^{p \times p}$
2: $n_{+1} \leftarrow |\{i \,:\, y_i = +1\}|$
3: $n_{-1} \leftarrow |\{i \,:\, y_i = -1\}|$
4: **for** $i = 1$ to n **do**
5: Compute Π_{x_i} (by sorting x_i)
6: $M_{LDA} \leftarrow M_{LDA} + \frac{y_i}{n_{y_i}} \Pi_{x_i}$
7: **end for**
8: $(\sigma, w, f) \leftarrow SVD(M_{LDA}, 1)$

- Ridge penalty (no monotonicity constraint), equivalent to rank-1 regression problem
- SVD finds the closest rank-1 matrix to the LDA solution:

$$M_{LDA} = \frac{1}{n_+} \sum_{i\,:\,y_i=+1} \Pi_{x_i} - \frac{1}{n_-} \sum_{i\,:\,y_i=+1} \Pi_{x_i}$$

- Complexity $O(np\ln(p))$ (same as QN only)

Experiments: Simulations

- True distribution of X entries is normal
- Corrupt data with a cauchy, exponential, uniform or bimodal gaussian distributions.
- $p = 1000$, n varies, logistic regression.