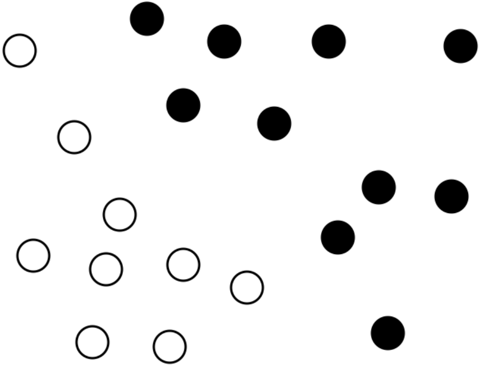
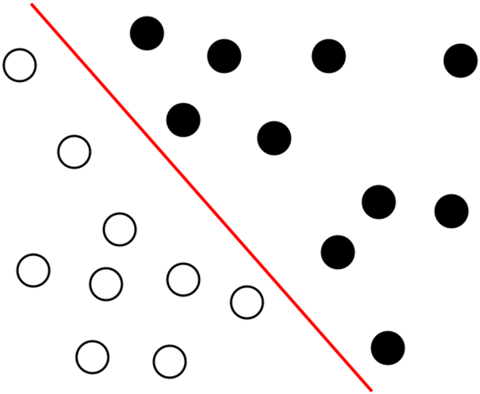
Machine learning on the symmetric group

Jean-Philippe Vert

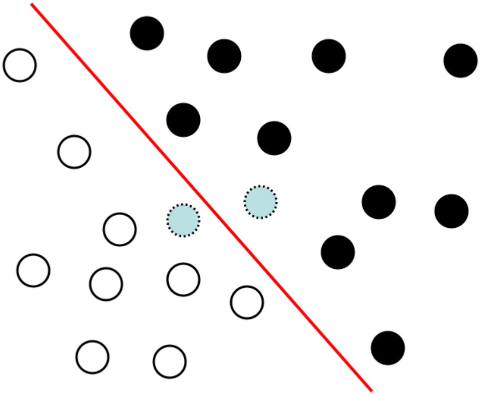
ML



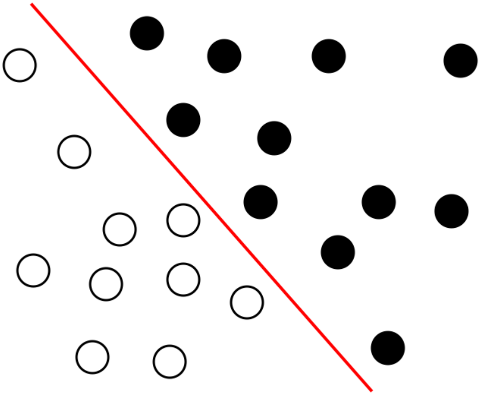
ML



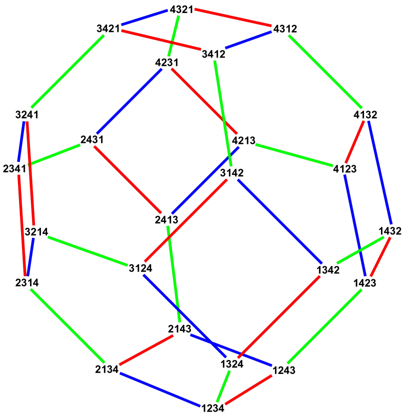
ML



ML



What if inputs are permutations?



- Permutation: a bijection

$$\sigma : [1, N] \to [1, N]$$

- $\sigma(i)$ = rank of item i
- Composition

$$(\sigma_1 \sigma_2)(i) = \sigma_1(\sigma_2(i))$$

- $\mathbb{S}_N$ the symmetric group
- $|\mathbb{S}_N| = N!$

Examples

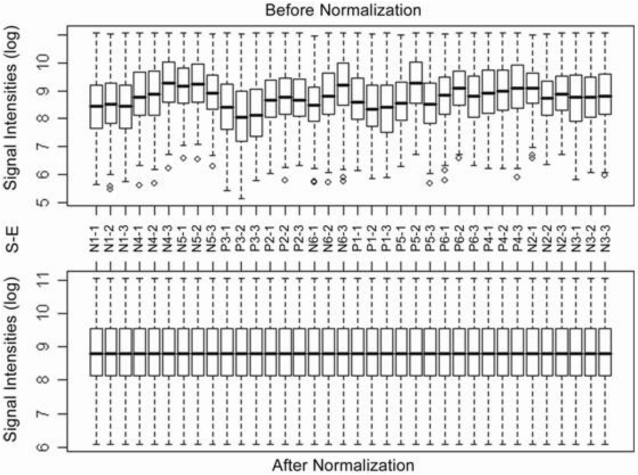
- Ranking data

- Ranks extracted from data

(histogram equalization, quantile normalization...)

Examples

- Batch effects, calibration of experimental measures



Learning from permutations

- Assume your data are permutations and you want to learn

$$f : \mathbb{S}_N \to \mathbb{R}$$

- A solutions: embed $\mathbb{S}_N$ to a Euclidean (or Hilbert) space

$$\Phi : \mathbb{S}_N \to \mathbb{R}^p$$

and learn a linear function:

$$f_\beta(\sigma) = \beta^\top \Phi(\sigma)$$

- The corresponding kernel is

$$K(\sigma_1, \sigma_2) = \Phi(\sigma_1)^\top \Phi(\sigma_2)$$

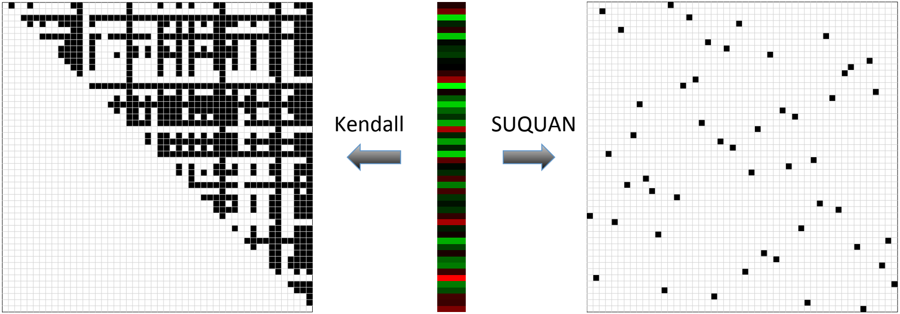
How to define the embedding $\Phi : \mathbb{S}_N \to \mathbb{R}^p$?

- Should encode interesting features
- Should lead to efficient algorithms

- Should be invariant to renaming of the items, i.e., the kernel should be right-invariant

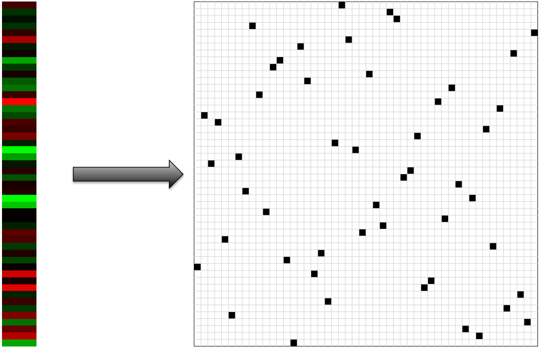
$$\forall \sigma_1, \sigma_2, \pi \in \mathbb{S}_N\,, \quad K(\sigma_1\pi, \sigma_2\pi) = K(\sigma_1, \sigma_2)$$

Some attempts



(Jiao and Vert, 2015, 2017, 2018; Le Morvan and Vert, 2017)

SUQUAN embedding (Le Morvan and Vert, 2017)



- Let $\Phi(\sigma) = \Pi_\sigma$ the permutation representation (Serres, 1977):

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

- Right invariant:

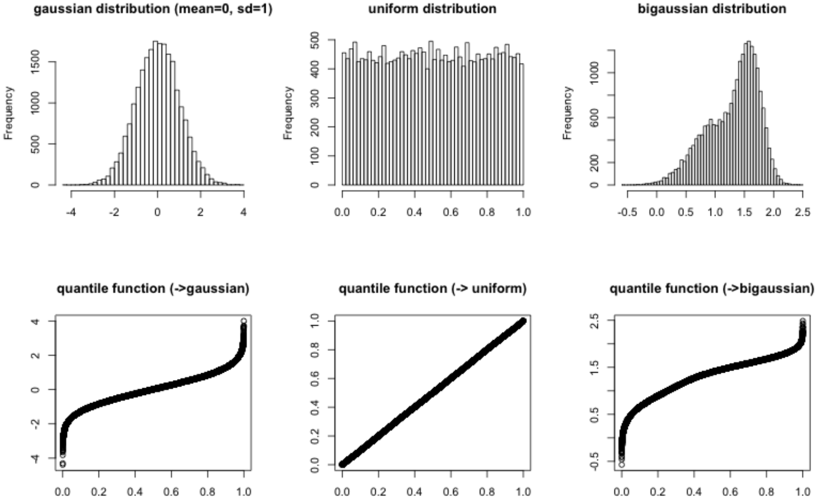
$$< \Phi(\sigma), \Phi(\sigma') > = \mathrm{Tr}\left(\Pi_\sigma \Pi_{\sigma'}^\top\right) = \mathrm{Tr}\left(\Pi_\sigma \Pi_{\sigma'}^{-1}\right) = \mathrm{Tr}\left(\Pi_\sigma \Pi_{\sigma'^{-1}}\right) = \mathrm{Tr}\left(\Pi_{\sigma\sigma'^{-1}}\right)$$

Link with quantile normalization (QN)

- Take $\sigma(x) = \texttt{rank}(x)$ with $x \in \mathbb{R}^N$
- Fix a target quantile $f \in \mathbb{R}^n$
- "Keep the order of x, change the values to f"

$$[\Psi_f(x)]_i = f_{\sigma(x)(i)} \quad \Leftrightarrow \quad \Psi_f(x) = \Pi_{\sigma(x)} f$$

How to choose a "good" target distribution?



Supervised QN (SUQUAN)

Standard QN:

1. Fix f arbitrarily
2. QN all samples to get $\Psi_f(x_1), \ldots, \Psi_f(x_N)$
3. Learn a model on normalized data, e.g.:

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

SUQUAN: jointly learn f and the model:

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

SUQAN as rank-1 matrix regression over $\Phi(\sigma)$

- Linear SUQAN therefore solves

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

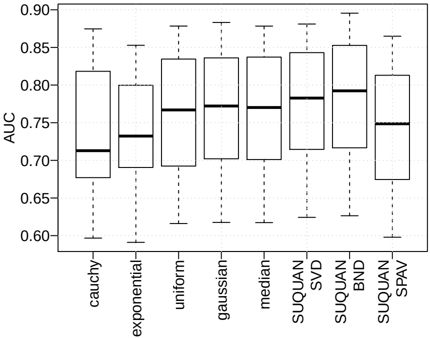
$$= \min_{w,b,f} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell \left(w^\top \Pi_{\sigma(x_i)}^\top f + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}$$

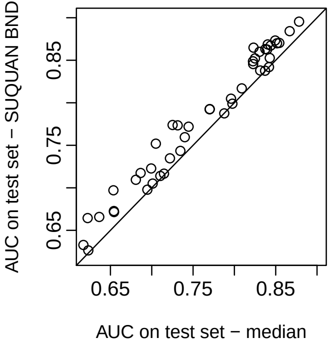
$$= \min_{w,b,f} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell \left(< \Pi_{\sigma(x_i)}, fw^\top >_{\text{Frobenius}} + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}$$

- 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

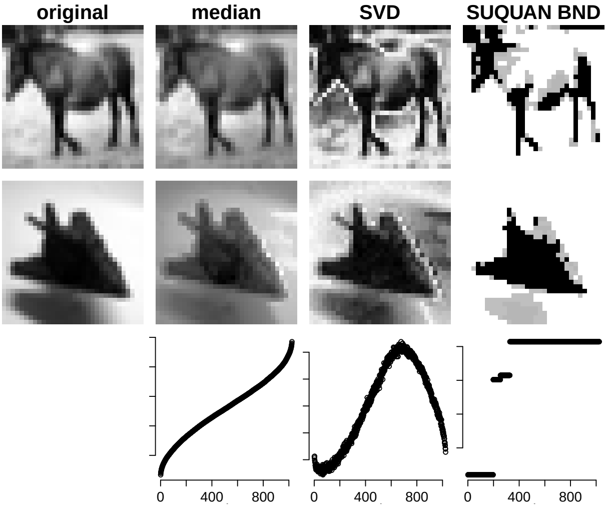
- Image classification into 10 classes (45 binary problems)
- $N = 5,000$ per class, $p = 1,024$ pixels
- Linear logistic regression on raw pixels



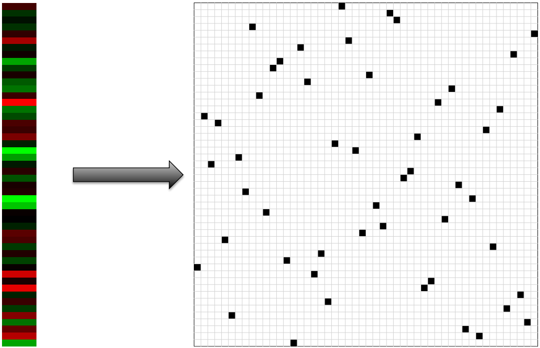


Experiments: CIFAR-10

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

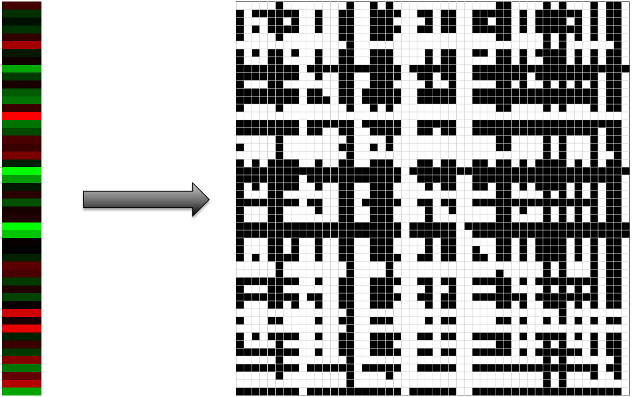


Limits of the SUQUAN embedding



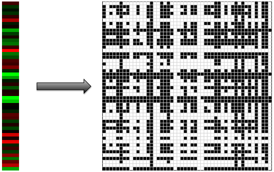
- Linear model on $\Phi(\sigma) = \Pi_\sigma \in \mathbb{R}^{N \times N}$
- Captures first-order information of the form "*i-th feature ranked at the j-th position*"
- What about higher-order information such as "*feature i larger than feature j*"?

The Kendall embedding (Jiao and Vert, 2015, 2017)



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

Geometry of the embedding



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

- Inner product

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

n_c = number of concordant pairs

- Distance

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

n_d = number of discordant pairs

Kendall and Mallows kernels

- The Kendall kernel is

$$K_\tau(\sigma, \sigma') = n_c(\sigma, \sigma')$$

- The Mallows kernel is

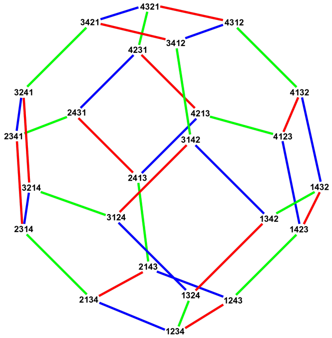
$$\forall \lambda \geq 0 \quad K_M^\lambda(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')}$$

Theorem (Jiao and Vert, 2015, 2017)

The Kendall and Mallows kernels are positive definite right-invariant kernels and can be evaluated in $O(N \log N)$ time

Kernel trick useful with few samples in large dimensions

Remark



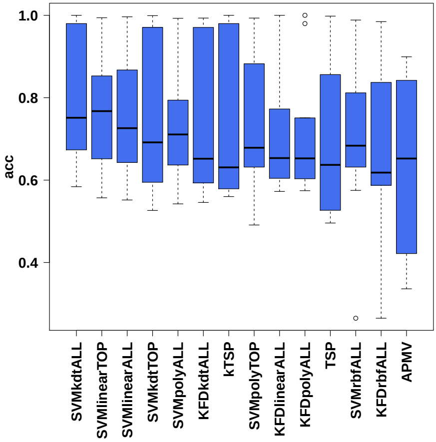
Cayley graph of $\mathbb{S}_4$

- 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

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

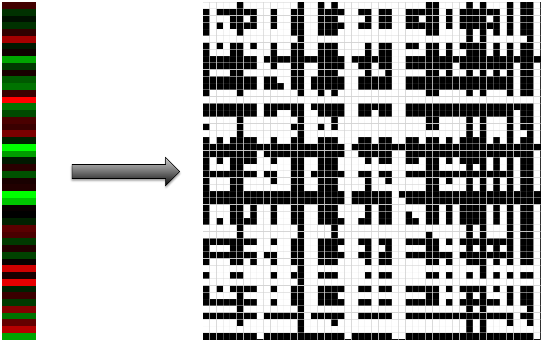
  where $n_d(\sigma, \sigma')$ is the shortest path distance on the Cayley graph.
- It can be computed in $O(N \log N)$

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:

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

Shieh (1998)

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

Kumar and Vassilvitskii (2010)

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

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

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

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

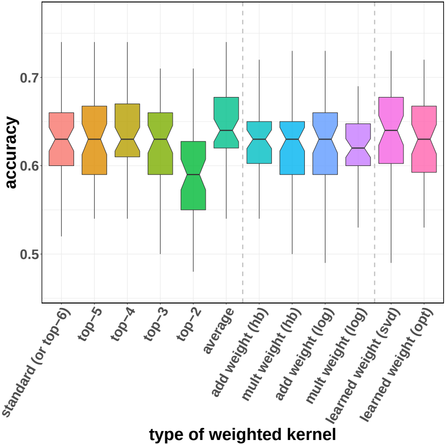
Learning the weights (2/2)

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

- 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
- Same as SUQAN, with $\Pi_\sigma \otimes \Pi_\sigma$ instead of Π_σ

Experiments

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



Towards higher-order representations

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

- 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:

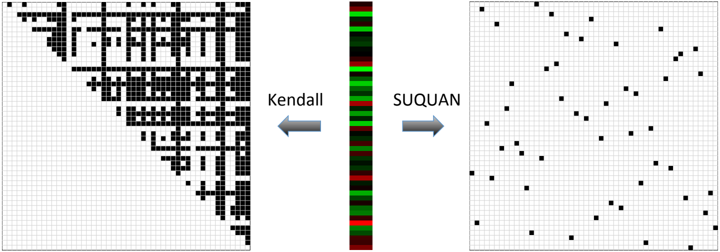
$$\Sigma \cong \tau_{(n-2,1,1)} \oplus \tau_{(n-1,1)}$$

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

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

- 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



- Machine learning beyond vectors, strings and graphs
- Different embeddings of the symmetric group
- Scalability? Robustness to adversarial attacks? Differentiable embeddings?

MERCI!

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.

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`.

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.

References (cont.)

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.

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.

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.

Harmonic analysis on $\mathbb{S}_N$

- A representation of $\mathbb{S}_N$ is a matrix-valued function $\rho : \mathbb{S}_N \to \mathbb{C}^{d_\rho \times d_\rho}$ such that
$$\forall \sigma_1, \sigma_2 \in \mathbb{S}_N\,, \quad \rho(\sigma_1\sigma_2) = \rho(\sigma_1)\rho(\sigma_2)$$
- A representation is irreductible (irrep) if it is not equivalent to the direct sum of two other representations
- $\mathbb{S}_N$ has a finite number of irreps $\{\rho_\lambda \,:\, \lambda \in \Lambda\}$ where $\Lambda = \{\lambda \vdash N\}$[1] is the set of partitions of N
- For any $f : \mathbb{S}_N \to \mathbb{R}$, the Fourier transform of f is

$$\forall \lambda \in \Lambda\,, \quad \hat{f}(\rho_\lambda) = \sum_{\sigma \in \mathbb{S}_N} f(\sigma)\rho_\lambda(\sigma)$$

[1] $\lambda \vdash N$ iff $\lambda = (\lambda_1, \dots, \lambda_r)$ with $\lambda_1 \geq \dots \geq \lambda_r$ and $\sum_{i=1}^r \lambda_i = N$

Right-invariant kernels

Bochner's theorem

An embedding $\Phi : \mathbb{S}_N \to \mathbb{R}^p$ defines a right-invariant kernel $K(\sigma_1, \sigma_2) = \Phi(\sigma_1)^\top \Phi(\sigma_2)$ if and only there exists $\phi : \mathbb{S}_N \to \mathbb{R}$ such that

$$\forall \sigma_1, \sigma_2 \in \mathbb{S}_N \,, \quad K(\sigma_1, \sigma_2) = \phi(\sigma_2^{-1}\sigma_1)$$

and

$$\forall \lambda \in \Lambda \,, \quad \hat{\phi}(\rho_\lambda) \succeq 0$$