Perm2vec

Jean-Philippe Vert

Tokyo Deep Learning Workshop, Tokyo, March 22, 2018

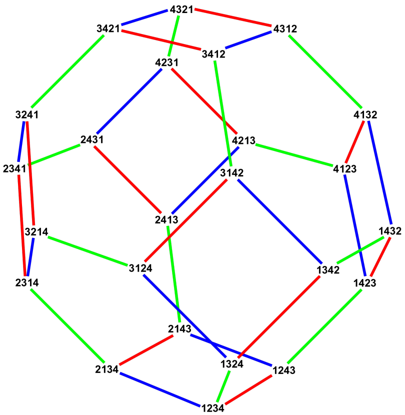
Motivations

- Ranking data

- Ranks extracted from data

(histogram equalization, quantile normalization...)

Mathematically



- Permutation: a bijection

$$\sigma : [1, n] \rightarrow [1, n]$$

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

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

- $\mathbb{S}_n$ the symmetric group
- $|\mathbb{S}_n| = n!$

Learning over the symmetric group

- 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

$$\Phi : \mathbb{S}_n \to \mathcal{H}$$

  and learn a function (e.g., linear):

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

- The corresponding kernel is

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

- A right-invariant kernel is invariant by renaming the items:

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

Outline

Outline

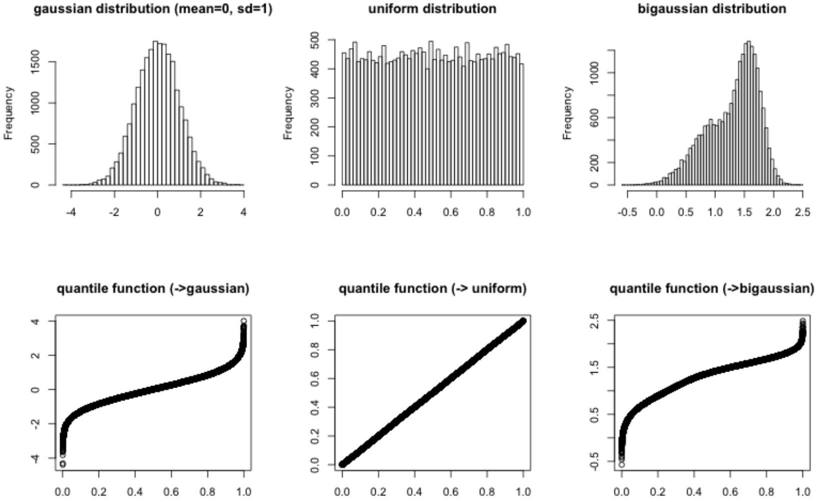
The quantile normalization (QN) embedding

- 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:

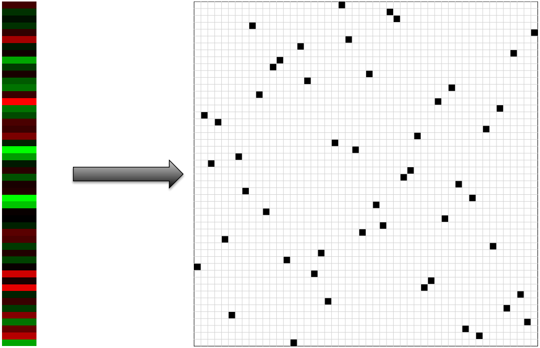
1. Fix f arbitrarily
2. QN all samples to get $\Phi_f(\sigma_1), \ldots, \Phi_f(\sigma_N)$
3. Learn a model on normalized data, e.g.:

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

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 SUQAN therefore solves

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

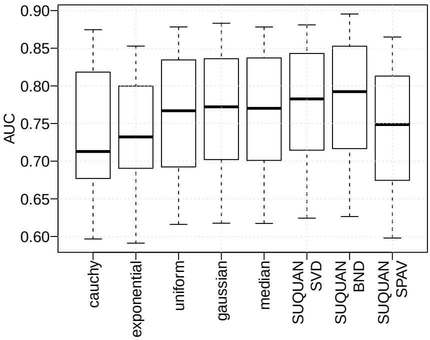
$$= \min_{w,b,f} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell \left(w^\top \Pi_{\sigma_i}^\top f + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}$$

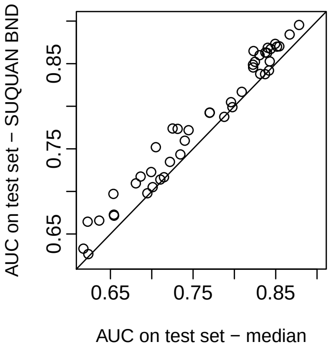
$$= \min_{w,b,f} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell \left(< \Pi_{\sigma_i}, fw^\top >_{\text{Frobenius}} + b \right) + \lambda \Omega(w) + \gamma \Omega_2(f) \right\}$$

- 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

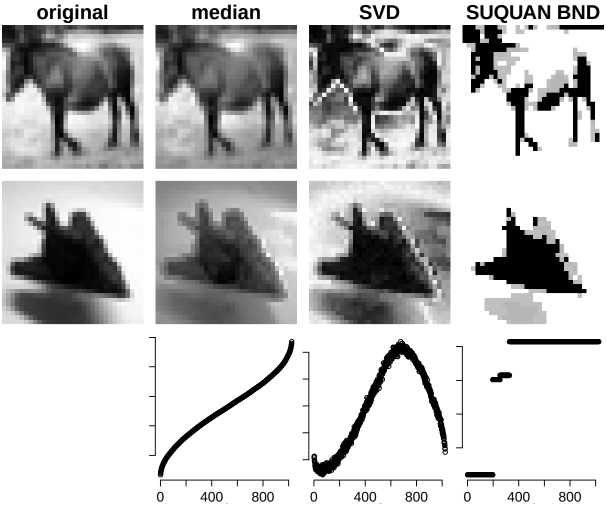
- Image classification into 10 classes (45 binary problems)
- $N = 5,000$ per class, $p = 1,024$ pixels





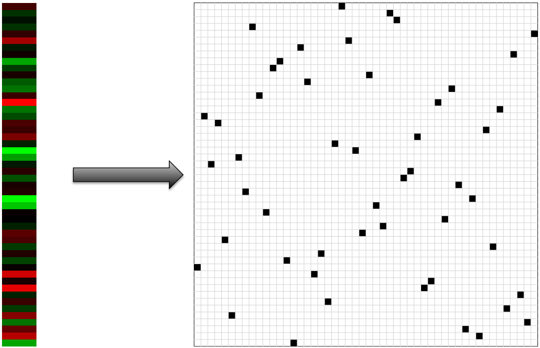
Experiments: CIFAR-10

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



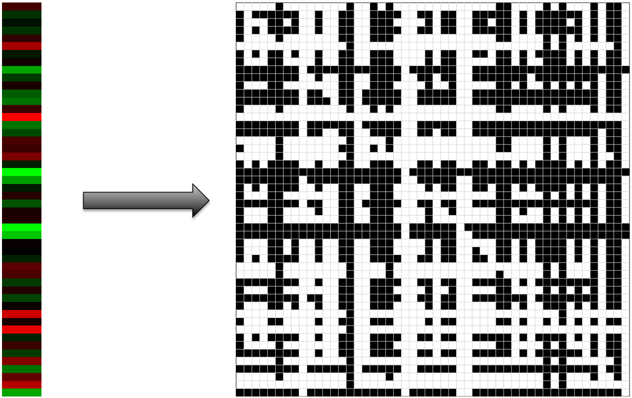
Outline

Limits of the QN embedding



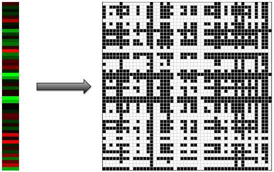
- Linear model on $\Phi(\sigma) = \Pi_\sigma \in \mathbb{R}^{n \times n}$
- Captures first-order information of the form "*i-th feature ranked at the j-th position*"
- What about higher-order information such as "*feature i larger than feature j*"?

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

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

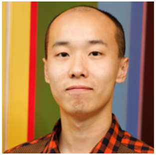
  n_c = number of concordant pairs

- Distance

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

  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)

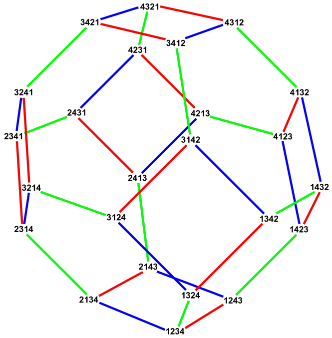
The Kendall and Mallows kernels are positive definite.

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



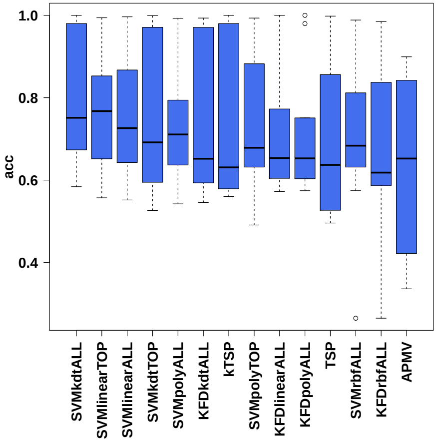
Cayley graph of $\mathbb{S}_4$

- 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

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

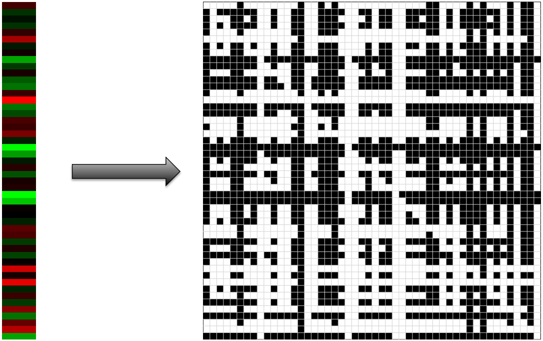
  where $n_d(\sigma, \sigma')$ is the shortest path distance on the Cayley graph.
- It can be computed in $O(n \log n)$

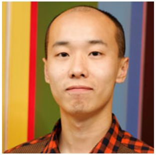
Applications



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:

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

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

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)

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

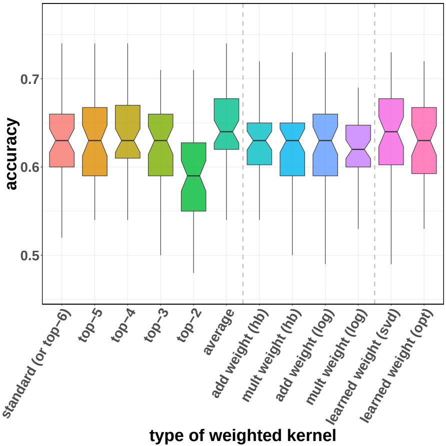
- This is symmetric in U and β
- Instead of fixing the weights U and optimizing β, we can jointly optimize β and U to learn the weights U
- Note that $\Pi_\sigma^\top = (\Pi_\sigma)^{-1} = \Pi_{\sigma^{-1}}$, hence

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

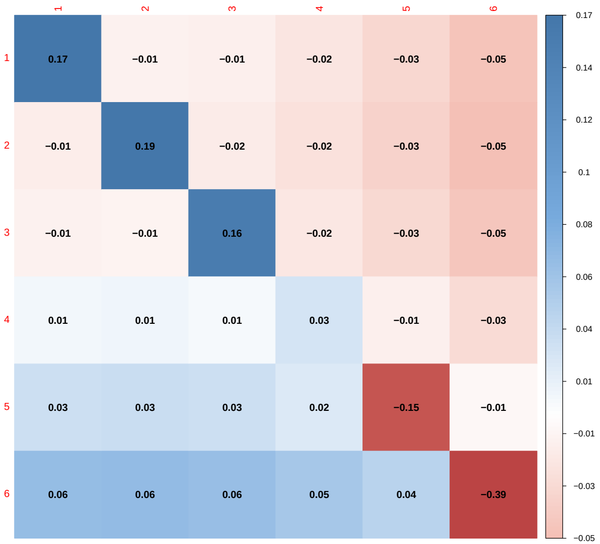
- We propose to alternate optimization in U and β
  - For U fixed, optimize β with $K_U(\sigma_1, \sigma_2)$
  - For β fixed, optimize U with $K_\beta(\sigma_1^{-1}, \sigma_2^{-1})$

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

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

- 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:

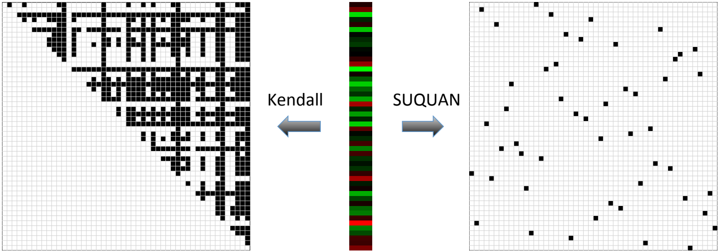
$$\Sigma \cong \tau_{(n-2,1,1)} \oplus \tau_{(n-1,1)}$$

- 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
- Respect the group structure (right-invariance) through group representations
- Compatible with NN architectures
- Scalability? Approximate embeddings?

Thanks

erc

SEVENTH FRAMEWORK PROGRAMME

INVESTISSEMENTS D'AVENIR

AGENCE NATIONALE DE LA RECHERCHE ANR

BIOMÉRIEUX

MINES ParisTech

institut Curie
Together, let's beat cancer.

Inserm
Institut national de la santé et de la recherche médicale

PSL RESEARCH UNIVERSITY

SANOFI

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868 LET THERE BE LIGHT

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

SIMONS INSTITUTE for the Theory of Computing

1794
ENS
ÉCOLE NORMALE SUPÉRIEURE

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.

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

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

- Non-decreasing and smooth

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

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.