The Weighted Kendall and High-order Kernels for Permutations

Yunlong Jiao[1], Jean-Philippe Vert[2]

[1]Department of Statistics & Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

[2]MINES ParisTech & Institut Curie & Ecole Normale Supérieure, PSL Research University, Paris, France

Overview

We study positive definite kernels for permutation/ranking data.

- They are weighted (and high-order) extensions of the Kendall kernel [1], allowing to weight differently the contributions of different items, e.g., to focus more on the top-ranked items.
- They are (symmetric,) positive definite, and invariant to shuffling of index of items to be ranked.
- They can be computed fast in $O(n\ln(n))$ operations.
- Weights can be learned systematically in a data-driven way.

Notations and Preliminaries

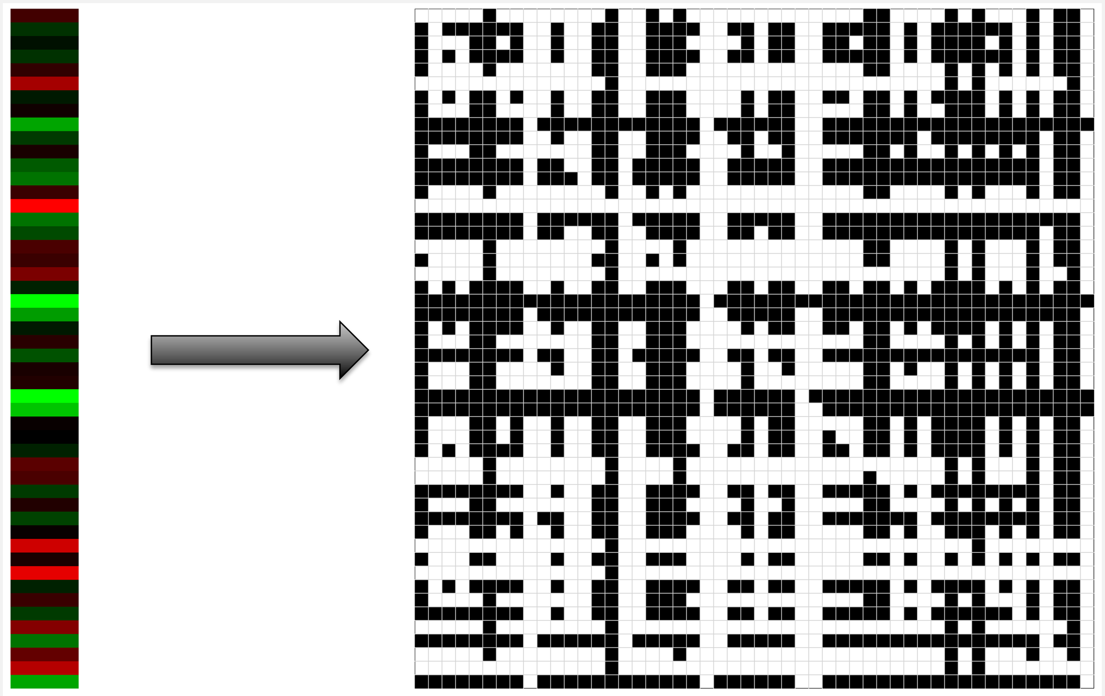
- A permutation σ is a 1-to-1 mapping of $[1,n]$ to itself.
- The symmetric group $\mathbb{S}_n$ is the set of all such permutations endowed with the composition operation.
- A positive definite (p.d.) kernel on $\mathbb{S}_n$ is a function $K:\mathbb{S}_n\times\mathbb{S}_n\to\mathbb{R}$ if there exists a Euclidean embedding $\Phi:\mathbb{S}_n\to\mathbb{R}^D$ such that
$$K(\sigma,\sigma')=\langle\Phi(\sigma),\Phi(\sigma')\rangle\,.$$
- A p.d. kernel K on $\mathbb{S}_n$ is right-invariant if for any σ,σ' and $\pi\in\mathbb{S}_n$, it holds that $K(\sigma,\sigma')=K(\sigma\pi,\sigma'\pi)$. A right-invariant kernel is invariant to shuffling of index of items to be ranked.

Background: The Kendall Kernel K_τ [1]

Take the Kendall embedding:
$$\Phi_\tau(\sigma)=\left(\mathbb{1}_{\sigma(i)<\sigma(j)}\right)_{1\le i,j\le n}\in\mathbb{R}^{n\times n}\,,$$
then the Kendall kernel is defined by the induced inner product:
$$K_\tau(\sigma,\sigma')=\langle\Phi_\tau(\sigma),\Phi_\tau(\sigma')\rangle=\sum_{1\le i,j\le n}\mathbb{1}_{\sigma(i)<\sigma(j)}\mathbb{1}_{\sigma'(i)<\sigma'(j)}\,.$$

Remark. The Kendall kernel K_τ amounts to the Kendall's τ correlation [3] up to constant shift and scaling (by taking $2K_\tau/\binom{n}{2}-1$).

Theorem (Kendall kernel [1, 2]). *K_τ is p.d., right-invariant, and can be computed in $O(n\ln(n))$ operations.*



Related Work: Weighted Kendall's τ

Given a weight function $w:[1,n]^2\to\mathbb{R}$, different weighted versions of the Kendall's τ correlation have been proposed:
$$\sum_{1\le i,j\le n}w(\sigma(i),\sigma(j))\mathbb{1}_{\sigma(i)<\sigma(j)}\mathbb{1}_{\sigma'(i)<\sigma'(j)}\ [4]$$
$$\sum_{1\le i,j\le n}\left(w(\sigma(i),\sigma(j))+w(\sigma'(i),\sigma'(j))\right)\mathbb{1}_{\sigma(i)<\sigma(j)}\mathbb{1}_{\sigma'(i)<\sigma'(j)}\ [5]$$
$$\sum_{1\le i,j\le n}w(\sigma(i),\sigma(j))\frac{p_{\sigma(i)}-p_{\sigma'(i)}}{\sigma(i)-\sigma'(i)}\frac{p_{\sigma(j)}-p_{\sigma'(j)}}{\sigma(j)-\sigma'(j)}\mathbb{1}_{\sigma(i)<\sigma(j)}\mathbb{1}_{\sigma'(i)<\sigma'(j)}\ [6]$$
Note that [4] reduces to the average precision correlation coefficient [7] by taking hyperbolic rank discounts $w(i,j)=1/(j-1)$. However, these functions are either not symmetric (hence not p.d.) [4, 6], or not p.d. [5].

The Weighted Kendall Kernel

Given a weight matrix $U\in\mathbb{R}^{n\times n}$, take the weighted Kendall embedding:
$$\Phi^U(\sigma)=\left(U_{\sigma(i),\sigma(j)}\mathbb{1}_{\sigma(i)<\sigma(j)}\right)_{1\le i,j\le n}\in\mathbb{R}^{n\times n}\,,$$
then the weighted Kendall kernel reduces to
$$K_U(\sigma,\sigma')=\langle\Phi^U(\sigma),\Phi^U(\sigma')\rangle=\sum_{1\le i,j\le n}U_{\sigma(i),\sigma(j)}U_{\sigma'(i),\sigma'(j)}\mathbb{1}_{\sigma(i)<\sigma(j)}\mathbb{1}_{\sigma'(i)<\sigma'(j)}\,.$$

Remark. Interesting choices of U include:

- *Top-k:* $U_{a,b}=1$ iff $a,b\le k$, for rank threshold $k\in[1,n]$.
- *Additive:* $U_{ij}=u_i+u_j$, for rank discounts $u\in\mathbb{R}^n$.
- *Multiplicative:* $U_{ij}=u_iu_j$, for rank discounts $u\in\mathbb{R}^n$.

In general, a systematic way to constructing a right-invariant, p.d., weighted Kendall kernel is as follows:

Theorem (Weighted Kendall kernel). *Let $W:\mathbb{N}^2\times\mathbb{N}^2\to\mathbb{R}$ be a p.d. kernel on $\mathbb{N}^2$, then the function $K_W:\mathbb{S}_n\times\mathbb{S}_n\to\mathbb{R}$ defined by*
$$K_W(\sigma,\sigma')=\sum_{1\le i,j\le n}W\left((\sigma(i),\sigma(j)),(\sigma'(i),\sigma'(j))\right)\mathbb{1}_{\sigma(i)<\sigma(j)}\mathbb{1}_{\sigma'(i)<\sigma'(j)}$$
is a right-invariant p.d. kernel on $\mathbb{S}_n$. If W is rank-1, K_W reduces to K_U.

Remark. Interesting general choices of W include:

- *Average:* $W\left((\sigma(i),\sigma(j)),(\sigma'(i),\sigma'(j))\right)=\min\{\sigma(i),\sigma'(i)\}/n$.

Theorem (Kernel trick). *The weighted Kendall kernels can be computed in $O(n\ln(n))$ for top-k, additive, multiplicative, or average weights.*

References

[1] Jiao and Vert. "The Kendall and Mallows kernels for permutations." *IEEE TPAMI*, 2018.
[2] Knight. "A computer method for calculating Kendall's tau with ungrouped data." *JASA*, 1966.
[3] Kendall. "A new measure of rank correlation." *Biometrika*, 1938.
[4] Shieh. "A weighted Kendall's tau statistic." *Statistics & Probability Letters*, 1998.
[5] Vigna. "A weighted correlation index for rankings with ties." *WWW*, 2015.
[6] Kumar and Vassilvitskii. "Generalized distances between rankings." *WWW*, 2010.
[7] Yilmaz, et al. "A new rank correlation coefficient for information retrieval." *SIGIR*, 2008.
[8] Le Morvan and Vert. "Supervised quantile normalisation." *arXiv:1706.00244*, 2017.

Learning the Weights

How to choose the pairwise position weights $U_{a,b}$? We propose to optimize them in a data-driven way in a supervised context.

Lemma. *Let us define the weighted embedding and kernel by*
$$\Phi^U(\sigma)=\left(U_{\sigma(i),\sigma(j)}\right)_{1\le i,j\le n}\in\mathbb{R}^{n\times n}\,,$$
$$G_U(\sigma,\sigma')=\langle\Phi^U(\sigma),\Phi^U(\sigma')\rangle\,,$$
then G_U reduces to K_U when $U\in\mathbb{R}^{n\times n}$ is zero in diagonal and lower-triangular, or is skew-symmetric (up to constant shift and scaling).

Theorem (Learning the weights). *Let us consider linear functions over the embedding Φ^U with coefficients $B\in\mathbb{R}^{n\times n}$, we have*
$$h^{U,B}(\sigma):=\langle B,\Phi^U(\sigma)\rangle=\langle U,\Phi^B(\sigma^{-1})\rangle \qquad (1a)$$
$$=\left\langle\mathrm{vec}(U)\otimes(\mathrm{vec}(B))^\top,\Pi_\sigma\otimes\Pi_\sigma\right\rangle\,, \qquad (1b)$$
where $(\Pi_\sigma)_{ij}=\mathbb{1}_{i=\sigma(j)}$ is the permutation representation.

Remark. The weights U and coefficients B can be learned jointly by solving a non-convex optimization via

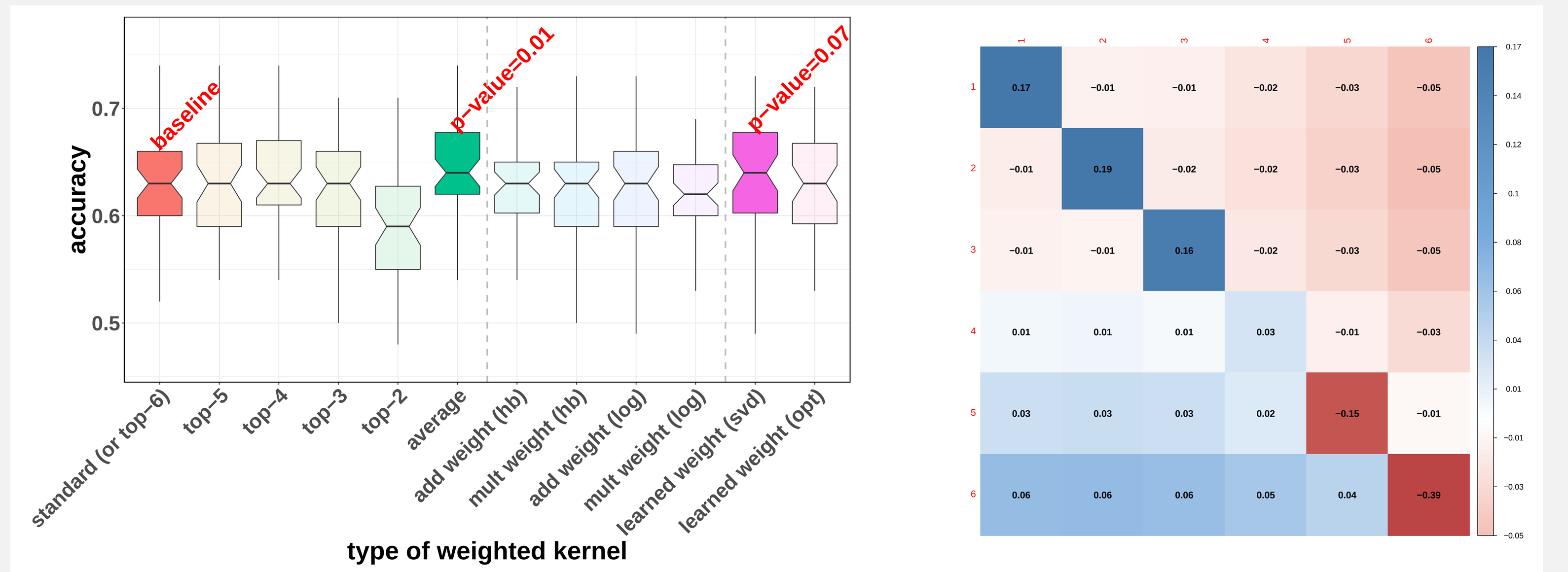
- Alternative optimization (1a).
- Low-rank approximation (1b), e.g., [8].

High-order Kernels

In order to consider three-way comparison (or higher-order in general), given a order-3 weight tensor $\mathcal{U}\in\mathbb{R}^{n\times n\times n}$, let us define the order-3 weighted embedding and kernel by
$$\Phi^{\mathcal{U}}(\sigma)=\left(\mathcal{U}_{\sigma(i),\sigma(j),\sigma(k)}\right)_{1\le i,j,k\le n}\in\mathbb{R}^{n\times n\times n}$$
$$G_{\mathcal{U}}(\sigma,\sigma')=\langle\Phi^{\mathcal{U}}(\sigma),\Phi^{\mathcal{U}}(\sigma')\rangle\,.$$
The three-way position weights $\mathcal{U}_{a,b,c}$ can also optimized in a data-driven way, due to the following results almost identical to the order-2 case.

Theorem (Learning the high-order weights). *Let us consider linear functions over the embedding $\Phi^{\mathcal{U}}$ with coefficients $\mathcal{B}\in\mathbb{R}^{n\times n\times n}$, we have*
$$h^{\mathcal{U},\mathcal{B}}(\sigma):=\langle\mathcal{B},\Phi^{\mathcal{U}}(\sigma)\rangle=\langle\mathcal{U},\Phi^{\mathcal{B}}(\sigma^{-1})\rangle=\langle\mathcal{U}\otimes\mathcal{B},\Pi_\sigma\otimes\Pi_\sigma\otimes\Pi_\sigma\rangle\,.$$

Numerical Experiments: Eurobarometer Survey Data



Left: Classification accuracy of predicting age group ($>/<$40yo) of >12k participants ranking the importance of $n=6$ sources of information. **Right:** Weights learned via low-rank approximation [8].