An equivalent formulation of matrix kernels (II)

Miaoyan Wang, Aug 3, 2020

1 Corrections to your section 1

Let $K(\cdot,\cdot)$ denote a usual kernel defined over vector pairs in $\mathbb{R}^d$. We use the shorthand $K(i,j) \stackrel{\text{def}}{=} K(\boldsymbol{X}_{i:}, \boldsymbol{X}'_{j:})$ to denote the kernel value evaluated on the vector pair $(\boldsymbol{X}_{i:}, \boldsymbol{X}'_{j:})$.

(Note: my definition of projection $\boldsymbol{P}$ seems to be the transpose of your definition of $\boldsymbol{P}$.)

Proposition 1 (Rank-1 weights in Kernel)**.** Define a kernel over matrix pairs, $\mathcal{K}(\boldsymbol{X}, \boldsymbol{X}') \stackrel{\text{def}}{=} \langle \boldsymbol{P}^T \Phi(\boldsymbol{X})\boldsymbol{P},\ \boldsymbol{P}^T\Phi(\boldsymbol{X}')\boldsymbol{P}\rangle$ for some rank-1 projection matrix $\boldsymbol{P} \in \mathbb{R}^{2\times 1}$. Then, $\mathcal{K}$ has an equivalent representation,

$$\mathcal{K}(\boldsymbol{X}, \boldsymbol{X}') = 2C \sum_{i,j} w_{ij} K(i,j), \tag{1}$$

where $\boldsymbol{W} = \boldsymbol{P}\boldsymbol{P}^T = [\![w_{ij}]\!]$ is a rank-1 weight matrix, and $C > 0$ is a normalizing constant.

Proof. By definition,

$$\begin{aligned}\mathcal{K}(\boldsymbol{X}, \boldsymbol{X}') &= \langle \boldsymbol{P}^T \Phi(\boldsymbol{X})\boldsymbol{P},\ \boldsymbol{P}^T\Phi(\boldsymbol{X}')\boldsymbol{P}\rangle \\ &= \langle \underbrace{\boldsymbol{P}\boldsymbol{P}^T}_{=:\boldsymbol{W}},\ \Phi^T(\boldsymbol{X}) \underbrace{\boldsymbol{P}\boldsymbol{P}^T}_{=:\boldsymbol{W}} \Phi(\boldsymbol{X}')\rangle \end{aligned} \tag{2}$$

Both $\boldsymbol{W}$ and $\Phi^T(\boldsymbol{X})\boldsymbol{W}\Phi(\boldsymbol{X}')$ are d-by-d matrices. The (i,j)-th entry of $\Phi^T(\boldsymbol{X})\boldsymbol{W}\Phi(\boldsymbol{X}')$ is

$$\begin{aligned}[\Phi^T(\boldsymbol{X})\boldsymbol{W}\Phi(\boldsymbol{X}')]_{ij} &= \sum_{s,s'} [\boldsymbol{\Phi}^T(\boldsymbol{X})]_{is}[\boldsymbol{W}]_{ss'}[\Phi(\boldsymbol{X}')]_{s'j} \\ &= \sum_{s,s'} w_{ss'} \big\langle (\phi(\boldsymbol{X}_{s:}), \phi(\boldsymbol{X}_{i:})),\ (\phi(\boldsymbol{X}'_{s':}), \phi(\boldsymbol{X}'_{j:}))\big\rangle \\ &= \sum_{s,s'} w_{ss'} \big(K(s,s') + K(i,j)\big) \\ &= CK(i,j) + \sum_{s,s'} w_{ss'}K(s,s'), \end{aligned} \tag{3}$$

where we have denoted the constant $C = \sum_{s,s'} w_{ss'} > 0$. Plugging (3) into (2) gives

$$\begin{aligned}\mathcal{K}(\boldsymbol{X}, \boldsymbol{X}') &= \sum_{i,j} w_{ij} [\Phi^T(\boldsymbol{X})\boldsymbol{W}\Phi(\boldsymbol{X}')]_{ij} \\ &= C\sum_{i,j} w_{ij}K(i,j) + \left(\sum_{i,j} w_{ij}\right)\left(\sum_{s,s'} w_{ss'}K(s,s')\right) \\ &= 2C\sum_{i,j} w_{ij}K(i,j). \end{aligned}$$

$\square$

Proposition 2 (Compatibility with row-wise-only mapping)**.** Based on your Section 2, the row-wise-only mapping induces the following kernel,

$$\langle \Phi(\boldsymbol{X})\boldsymbol{P},\ \Phi(\boldsymbol{X}')\boldsymbol{P}\rangle = \sum_{i,j} w_{ij}K(i,j),\ \text{where } \boldsymbol{W} = [\![w_{ij}]\!] = \boldsymbol{P}^T\boldsymbol{P} \text{ is a low-rank p.s.d. matrix.}$$

This kernel is proportional to that in (1).

2 Commentary to your section 3

Proposition 3 (Isomorphic Mappings; From Mapping to Kernel)**.** The following two mappings are isomorphic, in the sense that they induce the same kernel $\mathcal{K}$ over matrix pairs.

- Mapping 1

$$\begin{aligned} \Phi_1 : \mathbb{R}^{d_1\times d_2} &\to \mathcal{H}_r^{d_1} \times \mathcal{H}_c^{d_2} \\ \boldsymbol{X} &\mapsto (\Phi_r(\boldsymbol{X}), \Phi_c(\boldsymbol{X})) \stackrel{\text{def}}{=} (\phi_r(\boldsymbol{X}_{1:}), \ldots, \phi_r(\boldsymbol{X}_{d_1:}), \phi_c(\boldsymbol{X}_{:1}), \ldots, \phi_c(\boldsymbol{X}_{:d_2})) \end{aligned}$$

- Mapping 2

$$\begin{aligned} \Phi_2 : \mathbb{R}^{d_1\times d_2} &\to (\mathcal{H}_r \times \mathcal{H}_c)^{d_1\times d_2} \\ \boldsymbol{X} &\mapsto [\Phi_2(\boldsymbol{X})_{ij}], \quad \text{where } \Phi_2(\boldsymbol{X})_{ij} \stackrel{\text{def}}{=} (\phi_c(\boldsymbol{X}_{i:}),\ \phi_r(\boldsymbol{X}_{:j})) \end{aligned}$$

Proof. Using the similar argument in Proposition 1, we show that the kernel induced by (mapping 2 + low-rank coefficients) is

$$\begin{aligned} \mathcal{K}\colon \mathbb{R}^{d_1\times d_2} \times \mathbb{R}^{d_1\times d_2} &\to \mathbb{R} \\ \mathcal{K}(\boldsymbol{X}, \boldsymbol{X}') &\mapsto \sum_{i,j\in[d_1]} w_{ij}^{\text{row}} K_r(i,j) + \sum_{i,j\in[d_2]} w_{ij}^{\text{col}} K_c(i,j), \end{aligned} \tag{4}$$

where $\boldsymbol{W}^{\text{row}} = [\![w_{ij}^{\text{row}}]\!] = \frac{1}{c_1}\boldsymbol{P}_r\boldsymbol{P}_r^T$, $\boldsymbol{W}^{\text{col}} = [\![w_{ij}^{\text{col}}]\!] = \frac{1}{c_2}\boldsymbol{P}_c\boldsymbol{P}_c^T$ are some low-rank p.s.d. matrices, and $c_1 = \|\boldsymbol{1}_{d_1}^T\boldsymbol{P}_r\|_2^2 > 0$, $c_2 = \|\boldsymbol{1}_{d_2}^T\boldsymbol{P}_c\|_2^2 > 0$ are two normalizing constants.

Now, we consider the kernel induced by (mapping 1 + low-rank coefficients),

$$\begin{aligned} \mathcal{K}(\boldsymbol{X}, \boldsymbol{X}') &= \langle \Phi_r(\boldsymbol{X})\boldsymbol{P}_r,\ \Phi_r(\boldsymbol{X}')\boldsymbol{P}_r\rangle + \langle \Phi_c(\boldsymbol{X})\boldsymbol{P}_c,\ \Phi_c(\boldsymbol{X}')\boldsymbol{P}_c\rangle \\ &= \sum_{i,j\in[d_1]} w_{ij}^{\text{row}} K_r(i,j) + \sum_{i,j\in[d_2]} w_{ij}^{\text{col}} K_c(i,j), \end{aligned} \tag{5}$$

where $\boldsymbol{W}^{\text{row}} = [\![w_{ij}^{\text{row}}]\!]$, $\boldsymbol{W}^{\text{col}} = [\![w_{ij}^{\text{col}}]\!]$ are some low-rank p.s.d. matrices. $\square$

Two important properties in the induced kernels (4) and (5):

1. [Additivity] The new kernel is a linear combination of regular row and column kernels;
2. [Low-rank p.s.d.] The weight matrices $\boldsymbol{W}^{\text{row}}$, $\boldsymbol{W}^{\text{col}}$ are low-rank + p.s.d.

Conjecture 1 (From Kernel to Mapping)**.** Let $\mathcal{K}(\cdot,\cdot)$ be a function that maps a pair of matrices to a real-value. Suppose $\mathcal{K}(\cdot,\cdot)$ satisfies the above two properties. Then, the kernel $\mathcal{K}$ induces a decomposable feature mapping in that $\Phi(\boldsymbol{X}) = \Phi_r(\boldsymbol{X}) + \Phi_c(\boldsymbol{X})$, where, informally speaking, $\Phi_r(\cdot)$, $\Phi_c(\cdot)$ are the row- and column-wise mappings, respectively.

The decomposable mapping means the effects from rows and columns are additive/separable. Similar to an ANOVA model $Y_{ij} = \mu_i + \mu_j$ with marginal effects only. Additivity is a common assumption for matrix-based network analysis; see [**?**].