Entangled kernels

Riikka Huusari, Hachem Kadri

Qarma - LIS
Aix-Marseille University

IJCAI - 13.08.2019 - Macao

Main idea

How to learn non-separable operator-valued kernels?

Contributions:

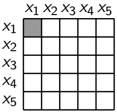
- Categorization of classes of operator-valued kernels
- Define two new classes of operator-valued kernels by leveraging tools from quantum computing.
- Derive algorithm for learning non-separable (entangled) operator-valued kernels.

Operator-valued Kernels

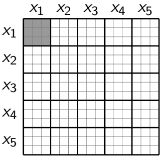
	Scalar-valued	Operator-valued
Kernel	$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{p \times p}$

Operator-valued Kernels

	Scalar-valued	Operator-valued
Kernel	$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{p \times p}$



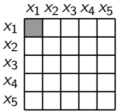
K



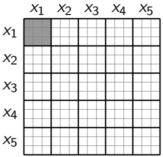
G

Operator-valued Kernels

	Scalar-valued	Operator-valued
Kernel	$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{p \times p}$



K



G

	Scalar-valued	Operator-valued
RKHS	$f : \mathcal{X} \to \mathcal{Y} \in \mathbb{R}$ $f \in \mathcal{K}$	$f : \mathcal{X} \to \mathcal{Y} \in \mathbb{R}^p$ $f \in \mathcal{H}$

Operator-valued Kernels

	Scalar-valued	Operator-valued
kernel trick	$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$	$\langle K(x, x')z, z' \rangle_{\mathcal{Y}} = \langle \phi(x)z, \phi(x')z' \rangle_{\mathcal{H}} \ \forall\, z, z' \in \mathcal{Y}$
representer theorem	$f(x) = \sum_i \alpha_i k(x_i, x)$ $\forall\, \alpha_i \in \mathbb{R}$	$f(x) = \sum_i K(x_i, x) c_i$ $\forall\, c_i \in \mathcal{Y}$

Some operator-valued kernels

Separable

$$K(\mathsf{x}, \mathsf{x}') = k(\mathsf{x}, \mathsf{x}')\mathbf{T} \qquad \mathbf{T} \in \mathcal{L}(\mathcal{Y})$$

Sum of separable

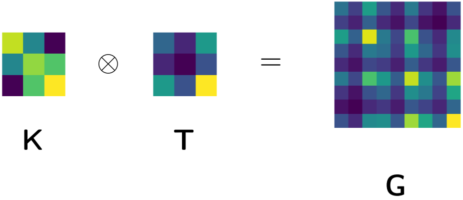
$$K(\mathsf{x}, \mathsf{x}') = \sum_i k_i(\mathsf{x}, \mathsf{x}')\mathbf{T}_i \qquad \mathbf{T}_i \in \mathcal{L}(\mathcal{Y})$$

Transformable

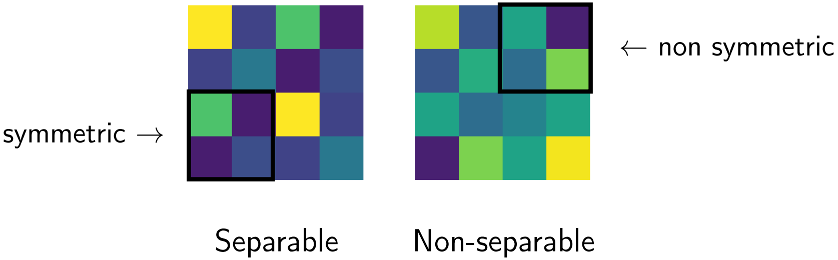
$$[K(\mathsf{x}, \mathsf{x}')]_{i,j} = \widetilde{k}(S_i\mathsf{x}, S_j\mathsf{x}') \qquad S_i, S_j \in \mathcal{L}(\mathcal{X}, \mathcal{H}_{\widetilde{k}})$$

Separable kernels

$$K(\mathbf{x}, \mathbf{z}) = k(\mathbf{x}, \mathbf{z})\mathbf{T}$$



Separable vs non-separable



Operator-valued kernel learning

- Multiple Operator-valued Kernel Learning: learn weights α_i from $K(\mathbf{x}, \mathbf{z}) = \sum_i \alpha_i k_i(\mathbf{x}, \mathbf{z}) \mathbf{T}_i$

Kadri H, Rakotomamonjy A, Preux P, Bach FR. Multiple operator-valued kernel learning. In Advances in Neural Information Processing Systems 2012.

Operator-valued kernel learning

- Multiple Operator-valued Kernel Learning: learn weights α_i from $K(\mathbf{x}, \mathbf{z}) = \sum_i \alpha_i k_i(\mathbf{x}, \mathbf{z}) \mathbf{T}_i$
- Learn $\mathbf{T}$ for separable operator-valued kernels
  - Takes only into account dependencies between output variables

Francesco Dinuzzo and Kenji Fukumizu. "Learning low-rank output kernels". In: Asian Conference on Machine Learning (ACML). 2011.

Operator-valued kernel learning

- Multiple Operator-valued Kernel Learning: learn weights α_i from $K(\mathbf{x}, \mathbf{z}) = \sum_i \alpha_i k_i(\mathbf{x}, \mathbf{z})\mathbf{T}_i$
- Learn $\mathbf{T}$ for separable operator-valued kernels
  - Takes only into account dependencies between output variables
- Learn Hadamard kernels: combination of separable and transformable kernels

Néhémy Lim et al. "Operator-valued kernel-based vector autoregressive models for network inference", Machine learning 2015

Operator-valued kernel learning

- Multiple Operator-valued Kernel Learning: learn weights α_i from $K(\mathbf{x}, \mathbf{z}) = \sum_i \alpha_i k_i(\mathbf{x}, \mathbf{z})\mathbf{T}_i$
- Learn $\mathbf{T}$ for separable operator-valued kernels
  - Takes only into account dependencies between output variables
- Learn Hadamard kernels: combination of separable and transformable kernels
- Multi-View Metric Learning: learn inseparable kernel for a multi-view problem

Huusari et al. "Multi-view Metric Learning in Vector-valued Kernel Spaces", AISTATS 2018

Operator-valued kernel learning

- Multiple Operator-valued Kernel Learning: learn weights α_i from $K(\mathbf{x}, \mathbf{z}) = \sum_i \alpha_i k_i(\mathbf{x}, \mathbf{z})\mathbf{T}_i$
- Learn $\mathbf{T}$ for separable operator-valued kernels
  - Takes only into account dependencies between output variables
- Learn Hadamard kernels: combination of separable and transformable kernels
- Multi-View Metric Learning: learn inseparable kernel for a multi-view problem

How to learn a general non-separable operator-valued kernel?

Classes of operator-valued kernels

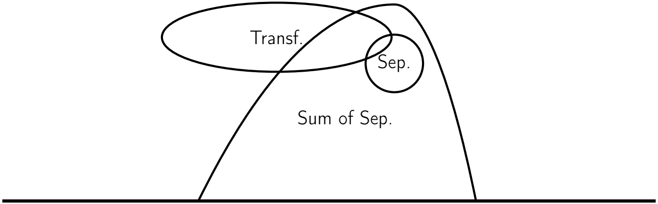


Illustration of inclusions among various operator-valued kernel classes.

Classes of operator-valued kernels

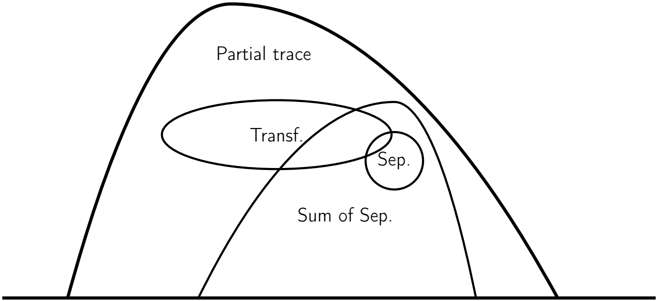


Illustration of inclusions among various operator-valued kernel classes.

Classes of operator-valued kernels

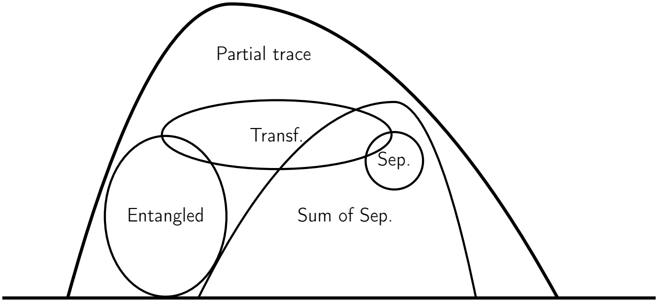


Illustration of inclusions among various operator-valued kernel classes.

Trace

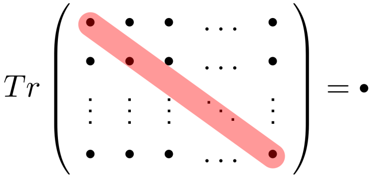


Illustration of trace operation.

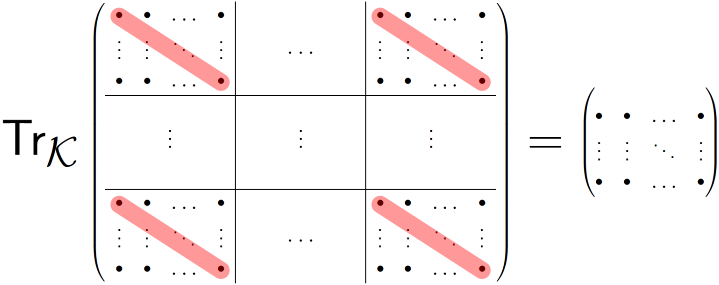


Illustration of partial trace operation. The partial trace operation applied to $N \times N$-blocks of a $Np \times Np$ matrix gives a $p \times p$ matrix as an output.

DEFINITION 1. (Partial trace kernel)
A partial trace kernel is defined as

$$K(\mathbf{x}, \mathbf{z}) = tr_{\mathcal{K}}(\mathbf{P}_{\phi(\mathbf{x}),\phi(\mathbf{z})}), \tag{1}$$

with $\mathbf{P}_{\mathbf{x},\mathbf{z}}$ is an operator on $\mathcal{L}(\mathcal{Y} \otimes \mathcal{K})$, and $tr_{\mathcal{K}}$ is the partial trace on $\mathcal{K}$ (i.e., over the inputs).

- Partial trace kernel generalizes the kernel trick:

$$k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = tr\big(\phi(\mathbf{x})\phi(\mathbf{z})^\top\big)$$

Definition 2. (Entangled kernel)

An entangled operator-valued kernel K is defined as

$$K(\mathbf{x}, \mathbf{z}) = tr_{\mathcal{K}}\left(\mathbf{U}\underbrace{\big(\underbrace{\mathbf{T}}_{p\times p}\otimes\underbrace{(\phi(\mathbf{x})\phi(\mathbf{z})^\top)}_{N\times N}\big)}_{pN\times pN}\mathbf{U}^\top\right), \tag{2}$$

where $\mathbf{U} \in \mathbb{R}^{pN\times pN}$ is not separable.

THEOREM 1. (Choi-Kraus representation)
The map $K(\mathbf{x},\mathbf{z}) = tr_{\mathcal{K}}\left(\mathbf{U}\big(\mathbf{T} \otimes (\phi(\mathbf{x})\phi(\mathbf{z})^\top)\big)\mathbf{U}^\top\right)$ can be generated by

$$K(\mathbf{x},\mathbf{z}) = \sum_{i=1}^{r} \mathbf{M}_i \phi(\mathbf{x})\phi(\mathbf{z})^\top \mathbf{M}_i^\top, \quad (3)$$

where $\mathbf{M}_i \in \mathbb{R}^{p \times N}$ and $1 \le r \le pN$.

$\Rightarrow$ with approximated feature map $\hat{\phi}(x) \in \mathbb{R}^m$ and simple linear algebra we have

$$\hat{\mathbf{G}} = (\hat{\mathbf{\Phi}}^\top \otimes \mathbf{I}_p)\mathbf{Q}\mathbf{Q}^\top(\hat{\mathbf{\Phi}} \otimes \mathbf{I}_p)$$

with $\mathbf{Q} \in \mathbb{R}^{mp*r}$

Kernel alignment:

$$A(\mathbf{M}, \mathbf{N}) = \frac{\langle \mathbf{M}_c, \mathbf{N}_c \rangle_F}{\|\mathbf{M}_c\|_F \|\mathbf{N}_c\|_F} \tag{4}$$

The optimization problem:

$$\max_{\mathbf{Q}} \quad (1-\gamma)\, A\left(tr_p(\hat{\mathbf{G}}), \mathbf{Y}^\top \mathbf{Y}\right) + \gamma\, A\left(\hat{\mathbf{G}}, \mathbf{y}\mathbf{y}^\top\right) \tag{5}$$

with $\gamma \in [0, 1]$.

Algorithm 1 Entangled Kernel Learning (EKL)

Input: matrix of features $\mathbf{\Phi}$, labels $\mathbf{Y}$
// 1) Kernel learning:
Solve for $\mathbf{Q}$ in eq.5 within a sphere manifold
// 2) Learning the predictive function:
if Predict with scalar-valued kernel **then**
  $\mathbf{c}_K = (tr_p(\hat{\mathbf{G}}) + \lambda \mathbf{I})^{-1}\mathbf{Y}^\top$ $\mathcal{O}(m^3 + mnp)$
else
  $\mathbf{c}_G = (\hat{\mathbf{G}} + \lambda \mathbf{I})^{-1}\,\text{vec}(\mathbf{Y})$ $\mathcal{O}(r^3 + mnp^2)$
Return $\mathbf{D} = \mathbf{Q}\mathbf{Q}^\top$, $\mathbf{c}$

Simulated data Simulated data created with bi-linear model $\mathbf{TCA} + \mathbf{ICK} = \mathbf{Y}$, algorithms are given $\mathbf{K}$ to solve for $\mathbf{Y}$.

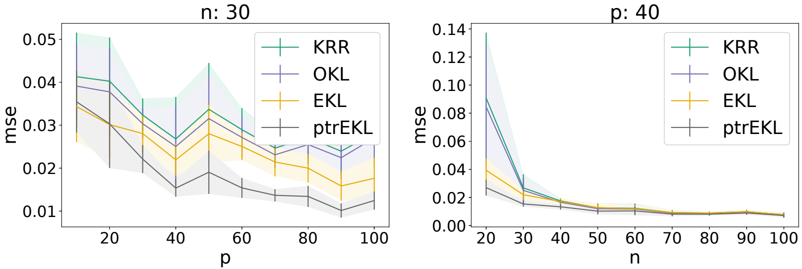


FIGURE: Results (mean squared error) of the simulated experiments.

EKL code available at
https://pageperso.lis-lab.fr/~riikka.huusari/research

Experiments

Real Weather-dataset with $p = 365$ and $n = 35$ (`www.psych.mcgill.ca/misc/fda`).

	$n = 5$		$n = 10$		$n = 15$	
method	nMSE	nl	nMSE	nl	nMSE	nl
KRR	0.951 ± 0.101	0.000	0.813 ± 0.141	0.000	0.761 ± 0.037	0.000
OKL	1.062 ± 0.250	-0.092	0.900 ± 0.196	-0.094	0.788 ± 0.058	-0.034
EKL/ptrEKL	0.840 ± 0.084	0.124	0.722 ± 0.036	0.107	0.728 ± 0.033	0.044

Table: Results on Weather data set averaged over 5 data partitions.

EKL code available at
https://pageperso.lis-lab.fr/~riikka.huusari/research

Conclusion

In this work we have

- defined a general class of kernels, partial trace kernels, that encompasses many OvK classes
- defined smaller class of entangled kernels that are not separable
- derived algorithm for learning entangled kernels in order to learn dependencies between input and output variables
- demonstrated the effectiveness of learning non-separable kernels
- connected the fields of quantum computing and machine learning by using notion of entanglement and the Choi-Kraus quantum separability theorem in context of kernel learning

Entangled Kernels

Riikka Huusari
riikka.huusari@lis-lab.fr

Hachem Kadri
hachem.kadri@lis-lab.fr

EKL code available at
https://pageperso.lis-lab.fr/riikka.huusari/research