خواجه نصیرالدین طوسی:
" هر کس چیزی را بجوید و در راهش کوشش نماید؛ آن را می یابد؛
هر کس دری را بکوبد و پایداری نماید؛ به درون خانه راه می یابد."

Shiraz University
Computer Science and Engineering Department
Machine Learning Lab

Positive Semi-Definite (PSD) Kernel Learning for Supervised and Unsupervised Problems

Dr. Sattar Hashemi

Fatemeh Alavi

alavi.s.fatemeh@gmail.com

MACHINE LEARNING
SMART SECURITY

1402/01/29

Outline

Introduction

PSD Kernels & Indefinite Kernels

PSD Kernel Learning

- Parametric Kernel Learning
- Data-Dependent Kernel Learning

Multiple Kernel Learning

- Classification and Clustering Tasks

Neighborhood Kernel Learning

- Classification and Clustering Tasks

Conclusion

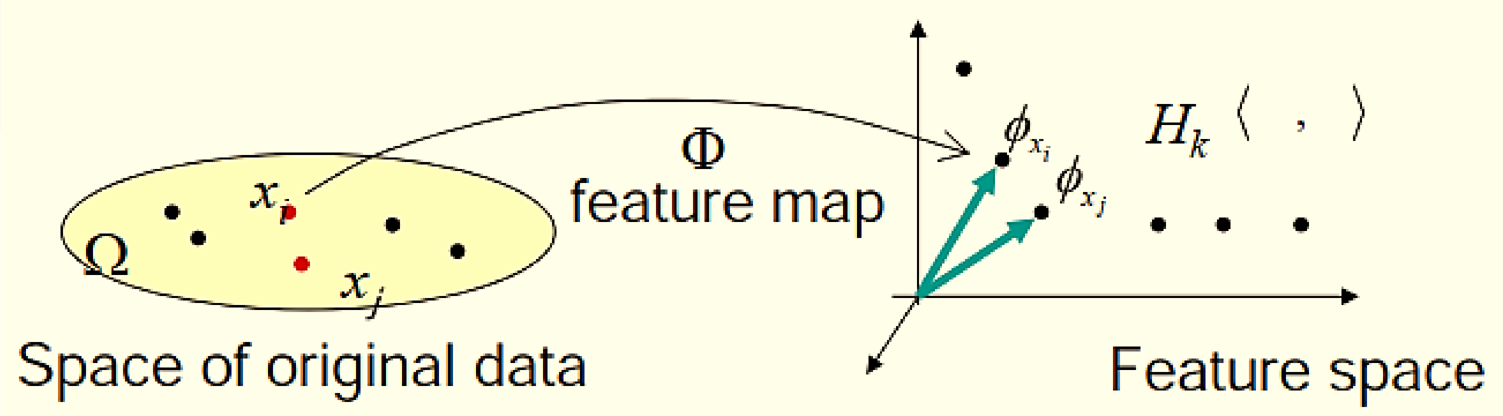
Introduction

- Traditional kernel methods **implicitly** embed data in some Hilbert spaces, and search for linear relations in the Hilbert spaces.
- The kernel matrix induces a notion of data **similarity** and data relationships.
- Kernels are considered as similarity measures that arise from a particular **representation** of patterns.

Kernel trick:

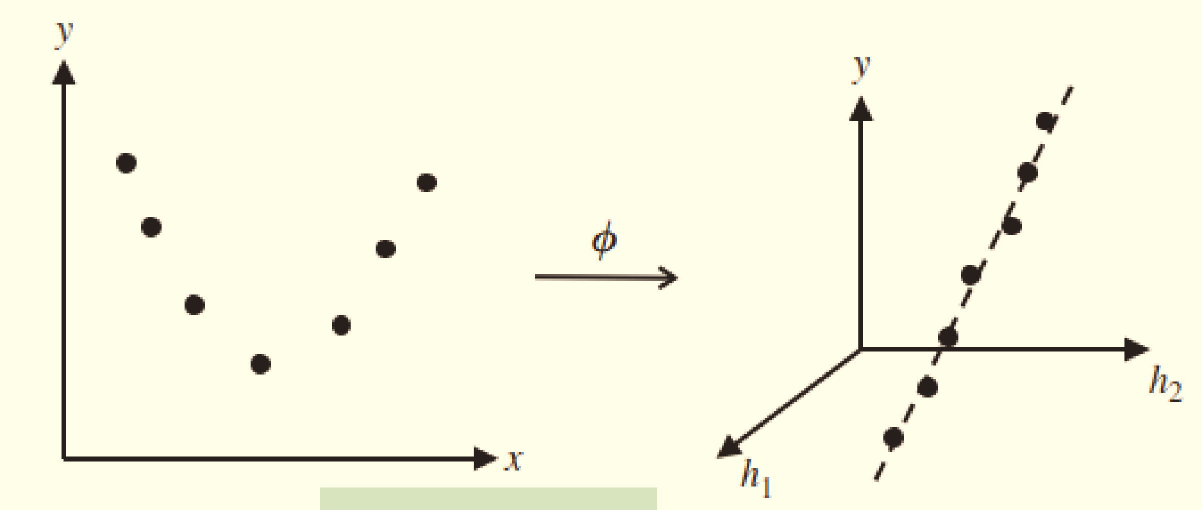
$$K(\boldsymbol{x}, \boldsymbol{x}') = < \emptyset(\boldsymbol{x}), \emptyset(\boldsymbol{x}') >_H$$

$$\text{Cos}(.) = \frac{<\emptyset(\boldsymbol{x}),\emptyset(\boldsymbol{x}')>_H}{\|x\|\|x\prime\|}$$

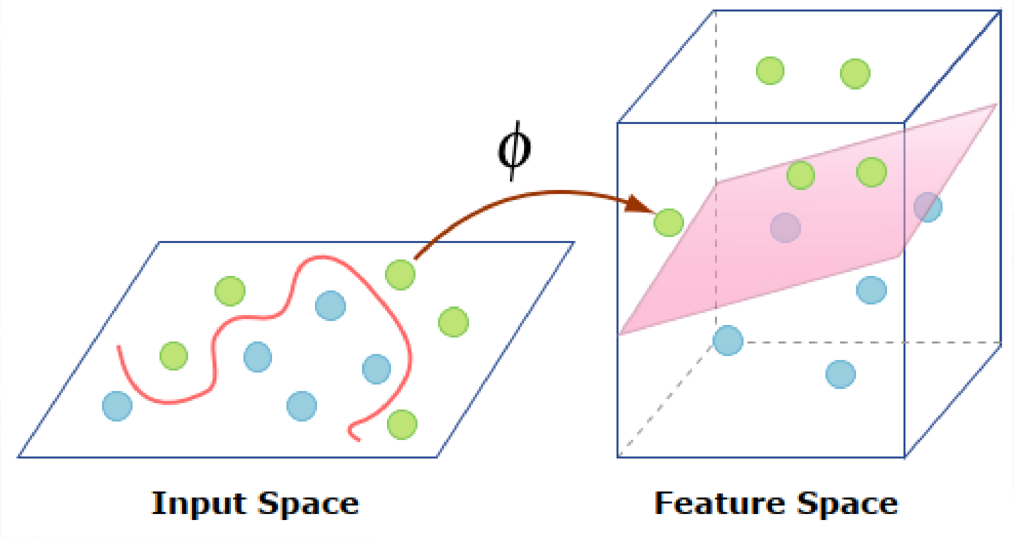


"Learning with kernels: support vector machines, regularization, optimization, and beyond." MIT press 2018.

Introduction (cont.)

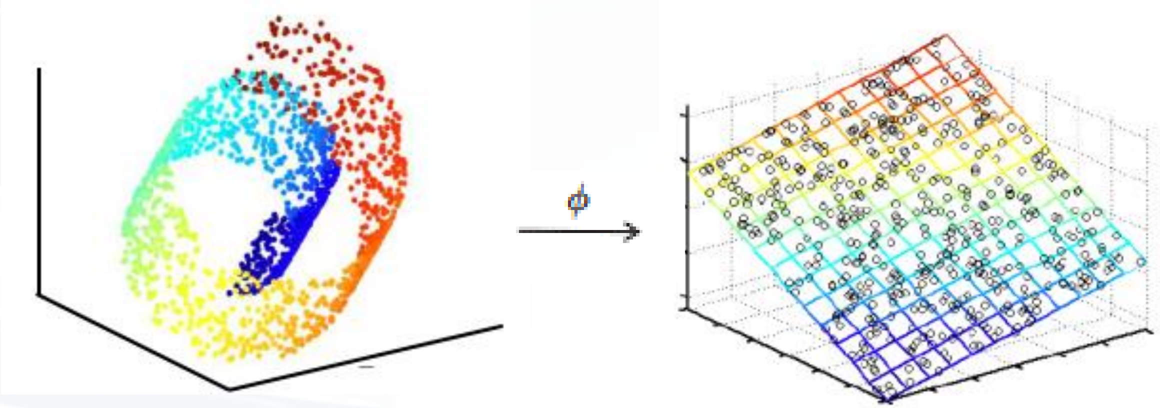


Regression



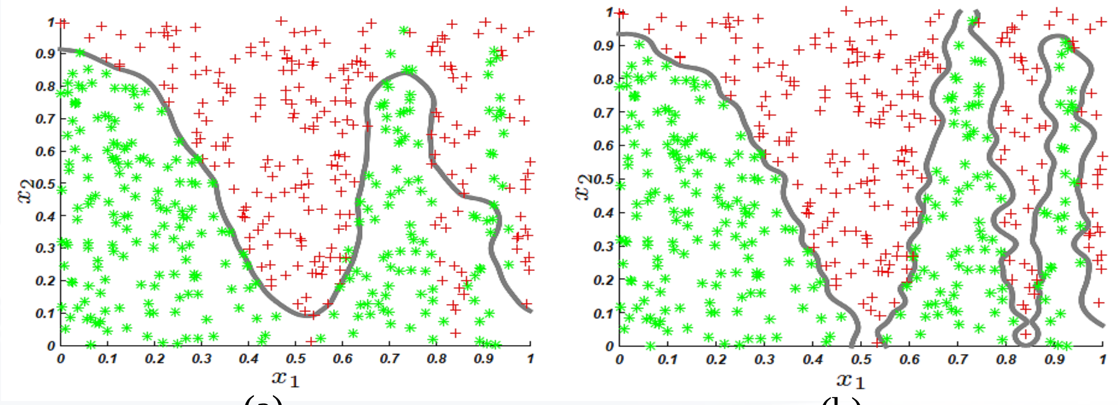
Classification

Unsupervised tasks



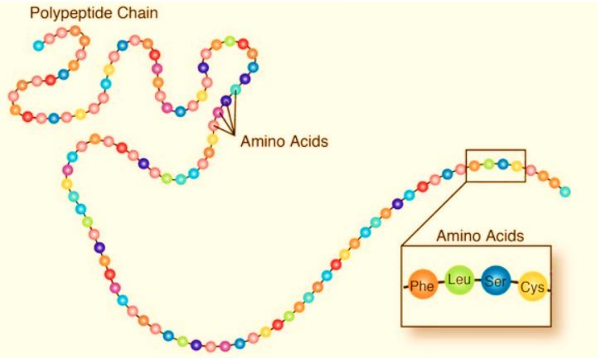
Introduction (cont.)

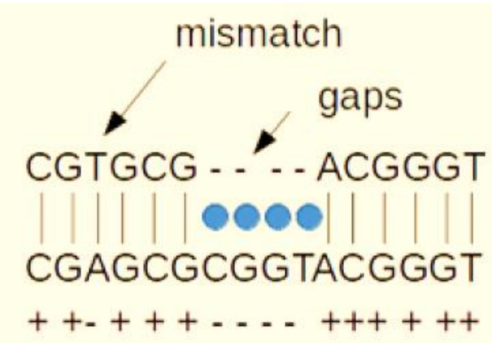
- The **performance** of kernel-based methods broadly relies on selecting an appropriate kernel.
- Traditional kernels **globally** perform on given data, and cannot **locally** capture the different relationships of data in different regions.
- Classical kernel matrixes are unable to capture **complex similarities**.



(a) (b)

Gaussian kernel: $e^{-\gamma\|x-x'\|_2^2}$ (a) $\gamma = 1$ (b) $\gamma = 25$

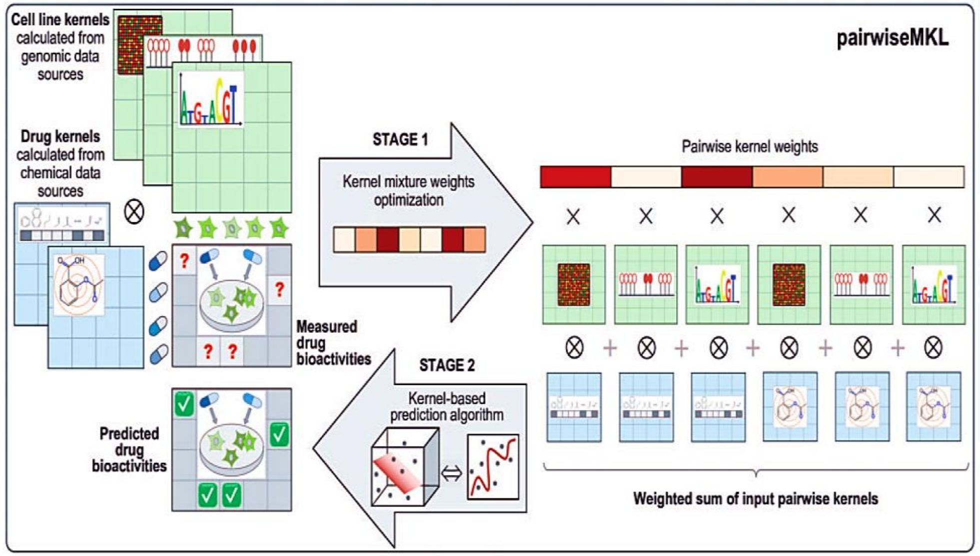




"Classification with Truncated ℓ1 Distance Kernel.", TPAMI 2015.

Applications of Kernel Learning

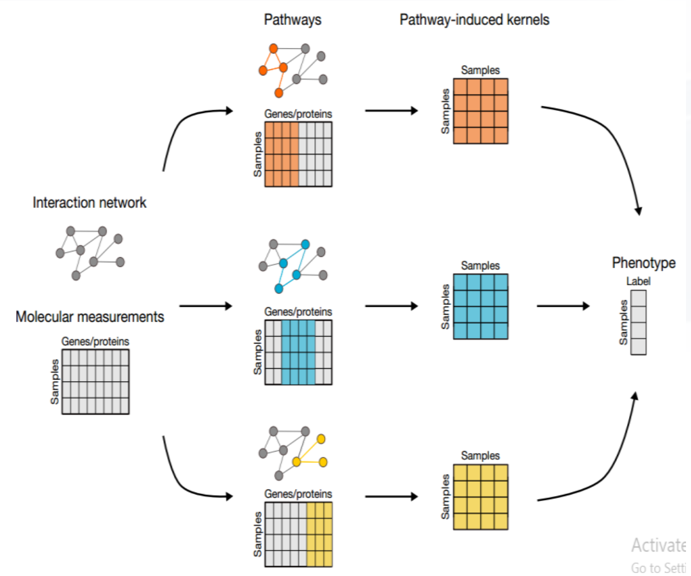
Drug Bioactivity Prediction



" Learning with multiple pairwise kernels for drug bioactivity prediction.", Bioinformatics 2018.

Bioinformatics: Patient Stratification

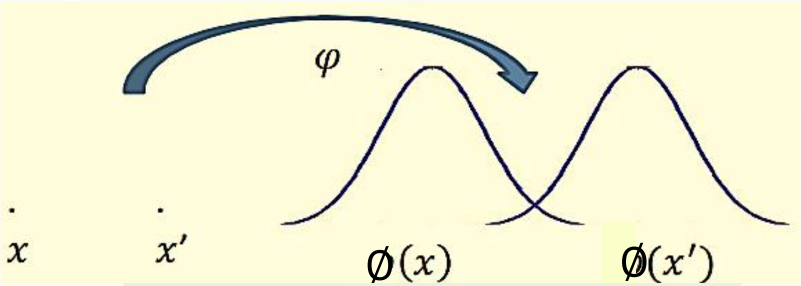
Pathway Induced Multiple Kernel Learning



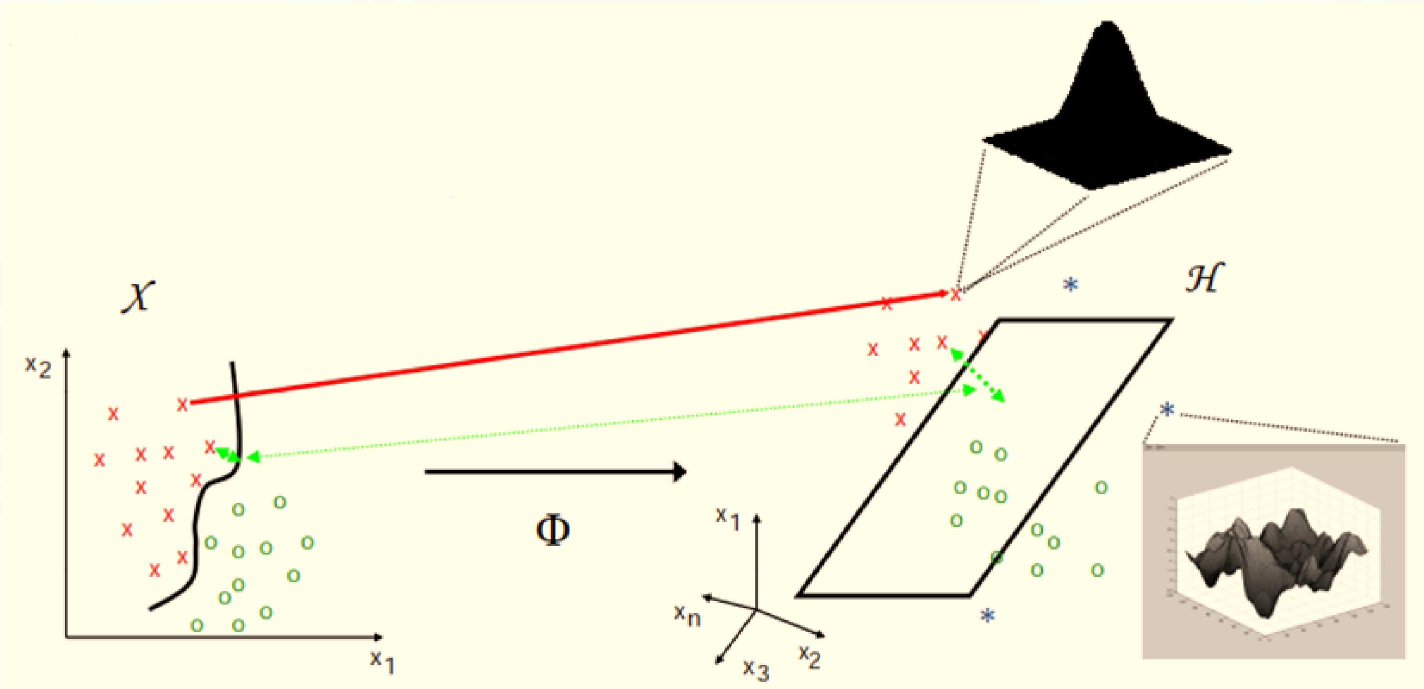
" PIMKL: Pathway-Induced Multiple Kernel Learning.", NPJ Systems Biology and Applications, 2019.

Deep Viewpoint

- **Bochner's theorem**: a kernel can be represented (in dual form) as a probability distribution, and so the search for a kernel becomes a search over distributions.



$$K(\boldsymbol{x}, \boldsymbol{x}') = < \emptyset(\boldsymbol{x}), \emptyset(\boldsymbol{x}') >_H$$



$$\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{K}(.,\boldsymbol{x}) = \boldsymbol{e}^{-\gamma\|.-\boldsymbol{x}\|^2}$$

$$\underbrace{\boldsymbol{\Phi}(\boldsymbol{x}') = \boldsymbol{K}(.,\boldsymbol{x}') = \boldsymbol{e}^{-\gamma\|.-\boldsymbol{x}'\|^2}}$$

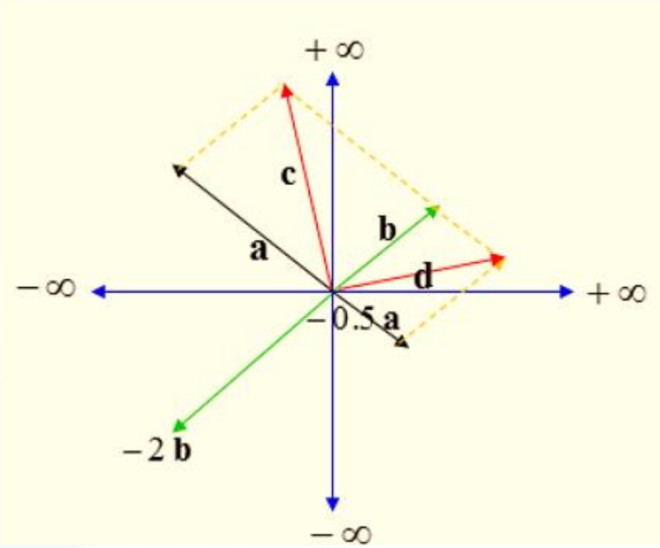
$$< \boldsymbol{\Phi}(\boldsymbol{x}), \boldsymbol{\Phi}(\boldsymbol{x}') > = K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{e}^{-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|^2}$$

"Learning with kernels: support vector machines, regularization, optimization, and beyond." MIT press 2018.

Hilbert Space (Functional Space)

- **Vector Space**:
  - V set of objects: vectors & functions
  - + vector addition: $\boldsymbol{a} \in V, \boldsymbol{b} \in V \ \rightarrow \boldsymbol{a} + \boldsymbol{b} \in V$
  - . scalar multiplication: $\mathbf{b} \in V, \alpha \in R \rightarrow \alpha\boldsymbol{b} \in V$

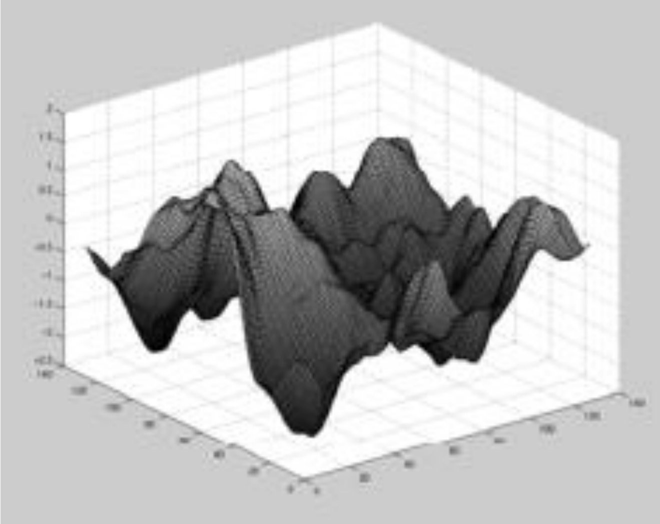
$$V = \{v | v = \sum_{i=1}^{n} \alpha_i \, \boldsymbol{x}_i \ \forall \boldsymbol{x}_i \in X\}$$



- **Hilbert space**: a complete vector space with dot product and a norm

- Functional space:

$$\boldsymbol{H} = \{f(.) | f(.) = \sum_{i=1}^{n} \alpha_i \, \boldsymbol{K}(., \boldsymbol{x}_i) = \sum_{i=1}^{n} \alpha_i \, \boldsymbol{\Phi}(\boldsymbol{x}_i) \ \ \forall \boldsymbol{x}_i \in X\}$$



Representer Theorem

Representer Theorem:

$$f^* = \underset{f(x)\in H}{\operatorname{argmin}}\, loss\big((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))\big) + \Omega(\|f\|_H{}^2)$$

although the optimization problem seems to be in an infinite-dimensional space, the solution only lies in the span of n particular kernels centered on n training points.

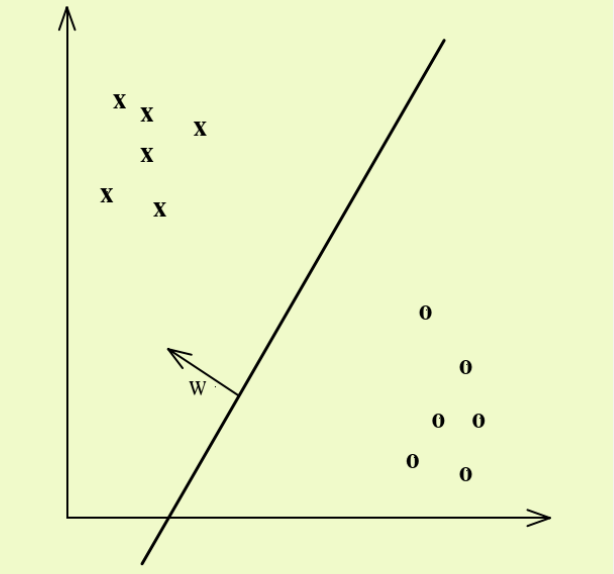
$$w = \sum_{i=1}^{n} \alpha_i\, \Phi(x_i) \rightarrow$$

$$f(x) =< w, \Phi(x) > +w_0 = \sum_{i=1}^{n} \alpha_i < \underbrace{\Phi(x_i), \Phi(x)}_{K(x_i, x)} > +w_0$$

$$w = \sum_{i=1}^{n} \alpha_i\, x_i \rightarrow$$

$$f(x) =< w, x_i > +w_0 = \sum_{i=1}^{n} \alpha_i < x_i, x > +w_0$$

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$
$$(x, x') \mapsto k(x, x'),$$



" Kernel methods for pattern analysis.", 2004, Cambridge university press.
"Learning with kernels: support vector machines, regularization, optimization, and beyond.", 2018, the MIT Press.
"Learning kernel classifiers: theory and algorithms.", 2001, MIT press.

Valid Kernel

Without constructing $\emptyset(\boldsymbol{x})$, Necessary and sufficient condition for k(x,x’) to be a kernel is:

Matrix kernel **K** is positive semi-definite (PSD)

Mercer’s theorem: any continuous, symmetric, positive semi-definite kernel function k(x,x’) can be expressed as a dot product in a high-dimensional space

$$\boldsymbol{K} = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}$$

- $k(x_i, x_i) \geq 0 \; \forall x_i \in X$
- All eigenvalues of $\boldsymbol{K}$ satisfy $\lambda_i \geq 0$

PSD Kernels vs. Indefinite Kernels

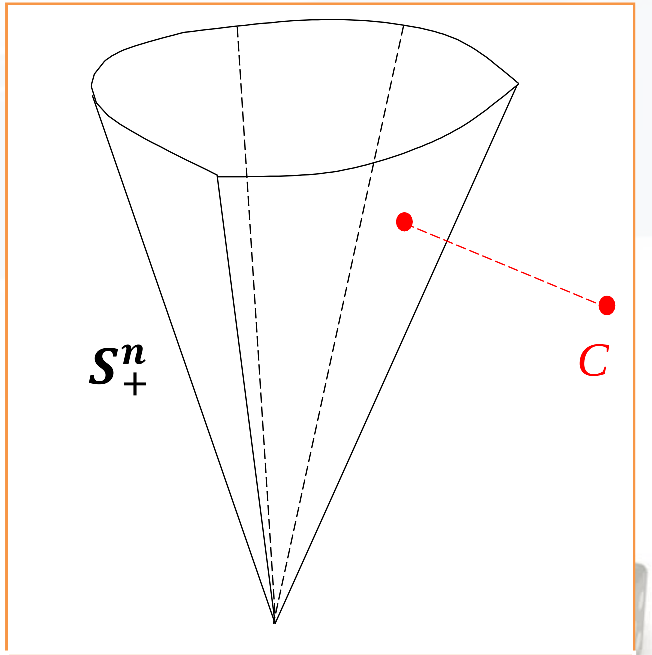
Indefinite Kernels
RKKS Space

- Representer theorem
- Non-convex problems
- No Mercer's condition
- No kernel trick

PSD Kernels
RKHS Space

- Representer theorem
- Convex problems
- Mercer's condition
- Kernel trick

S^n : the space of symmetric matrices
S^n_+ : cone of PSD matrices



Projection onto the PSD

"Classification with Truncated ℓ1 Distance Kernel.", TPAMI 2015.
"Indefinite Kernel Logistic Regression with Concave-Inexact-Convex Procedure.", TNNLS 2018.

PSD Kernel Learning Methods

Parametric Kernel Learning
(Inductive Learning)

- Multiple Kernel Learning
  - Linear & Nonlinear Kernel Learning
  - Finite & Infinite Kernel Learning
- Deep Kernel Learning
- Bayesian Multiple Kernel Learning

Data-dependent Kernel Learning
(Transductive Learning)

- Optimal Neighborhood Kernel Learning
- Low-rank Kernel Learning

Parametric Kernel Learning: Multiple KL

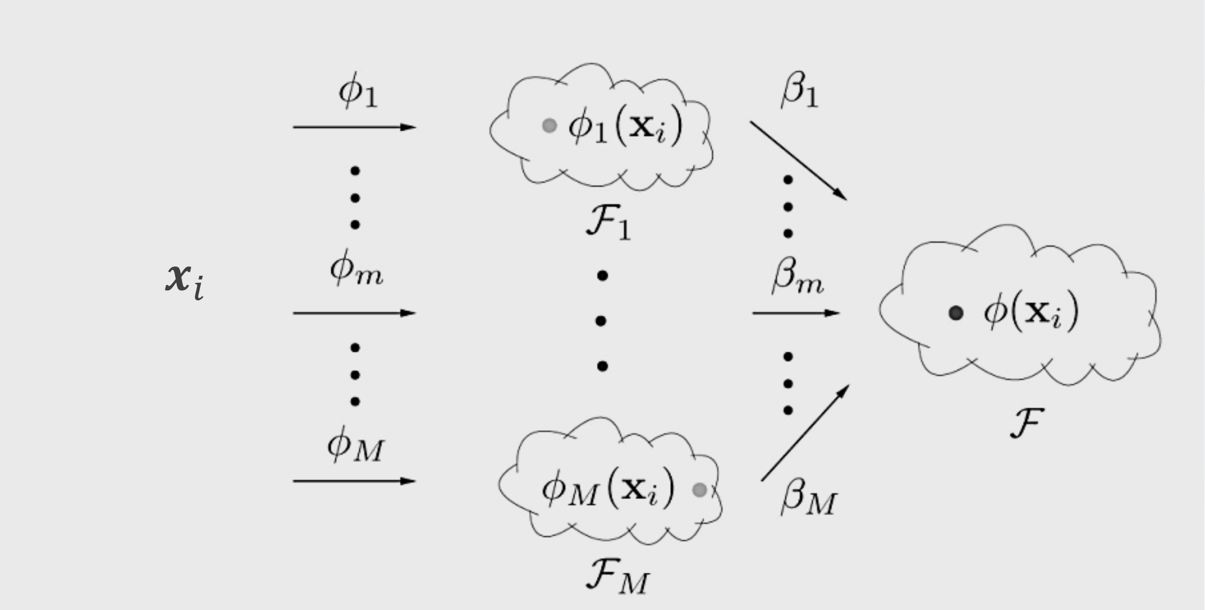
Finite kernel learning

- Base kernels should be specified **in advance**.
- Optimal kernel is a weighted combination of predefined base kernels over a parametrized **discrete** set.
- **Gaussian kernel**: $e^{-\gamma\|x-x'\|_2^2}$, $\gamma = 2^{-15}, \dots, 2^{15}$

Infinite kernel learning

$$K^{infinite} = \left\{\int_{\Omega} K_{\theta}dp(\theta): p \in M(\Omega)\right\}; \theta \in \Omega$$

- Base kernels can be learned **automatically**.
- Infinite kernel learning improves the accuracy.
- Optimal kernel is a weighted combination of learned base kernels over a **continuous** set.
- **Gaussian kernel**: $e^{-\mu\|x-x'\|_2^2}$, $\mu \in [2^{-15}, 2^{15}]$



Different representations

Different notions of similarity

"Multiple kernel learning algorithms.", JMLR 2011.
"Infinite kernel learning: generalization bounds and algorithms.", AAAI 2017.

Deep Kernel Learning

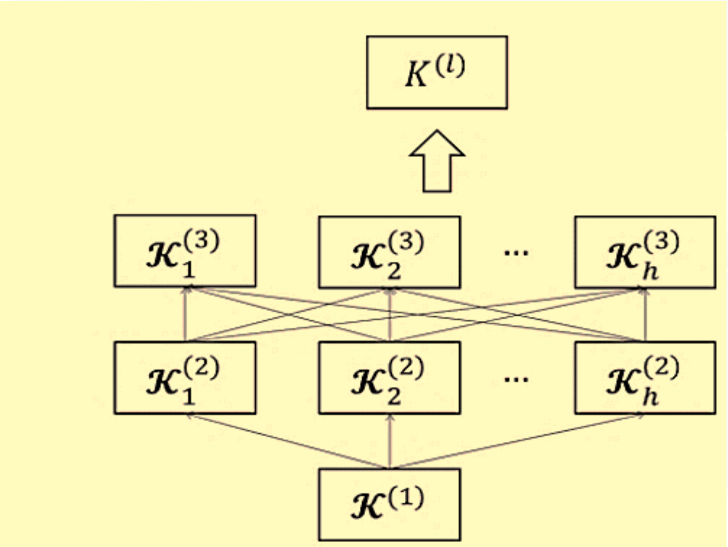
Multi-layer Composition of kernels

$$\mathcal{K}^{(l)}(\mathbf{x}_i, \mathbf{x}_j) = \phi^{(l)}\left(\cdots\phi^{(1)}(\mathbf{x}_i)\right) \cdot \phi^{(l)}\left(\cdots\phi^{(1)}(\mathbf{x}_j)\right)$$

$$\boldsymbol{K}^{(2)} = \emptyset^{(2)}\left(\emptyset^{(1)}(\boldsymbol{x}_i)\right).\emptyset^{(2)}\left(\emptyset^{(1)}(\boldsymbol{x}_j)\right) = e^{-2\gamma}e^{2\gamma K(x_i x_{j,})}$$

Nested Kernel:

$$\boldsymbol{K}^{(l)} = \{w_{1,1}^{(l)}\boldsymbol{K}_{1,1}^{(l)}\left(w_{1,1}^{(l-1)}\boldsymbol{K}_{1,1}^{(l-1)} + \cdots\right) + \cdots w_{h,m}^{(l)}\boldsymbol{K}_{h,m}^{(l)}(\ldots)\}$$



Deep Learning (Advantages):

- End-to-end learning
- "richness" of representations

Deep Learning (Limitations):

- Too many hyper-parameters
- a single type of kernel
- Tuning of several hyper-parameters in the discrete space

" Two-Layer Multiple Kernel Learning." , AISTATS 2011.
"A Representer Theorem for Deep Kernel Learning ." JMLR 2019.

MKL (Classification Tasks)

Kernel Target Alignment (KTA) shows that in the feature space the data distribution is somehow correlated to the label distribution.

$$KTA(\boldsymbol{K}_1, \boldsymbol{K}_2) = \frac{\langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F}{\|\boldsymbol{K}_1\|_F \|\boldsymbol{K}_2\|_F} \Rightarrow$$

Ideal Kernel:

$$\boldsymbol{K}_y = \boldsymbol{y}\boldsymbol{y}^T \begin{cases} K_{ij}^y = 1 \; if \; y_i = y_j \\ K_{ij}^y = -1 \; if \; y_i \neq y_j \end{cases}$$

"On kernel-target alignment. ", NIPS 2001.

Alignf Algorithm

$$\boldsymbol{a} = [< \boldsymbol{K}_1, \boldsymbol{K}_y >_F \quad \dots \quad < \boldsymbol{K}_m, \boldsymbol{K}_y >_F], \qquad M_{ij} =< \boldsymbol{K}_i, \boldsymbol{K}_j >_F = Tr(\boldsymbol{K}_i \boldsymbol{K}_j)$$

QP:

$$\max_{\boldsymbol{w}} \frac{< \boldsymbol{K}_w, \boldsymbol{K}_y >_F}{\sqrt{< \boldsymbol{K}_w, \boldsymbol{K}_w >_F}}, s.t. \; \|\boldsymbol{w}\|_2 = 1, \boldsymbol{w} \geq \boldsymbol{0}, \boldsymbol{K}_w = \sum_{p=1}^{m} w_p \boldsymbol{K}_p \rightarrow \min_{\|\boldsymbol{w}\|_2 = 1, \boldsymbol{w} \geq 0} \boldsymbol{w}^T \boldsymbol{M} \boldsymbol{w} - 2\boldsymbol{w}^T \boldsymbol{a}$$

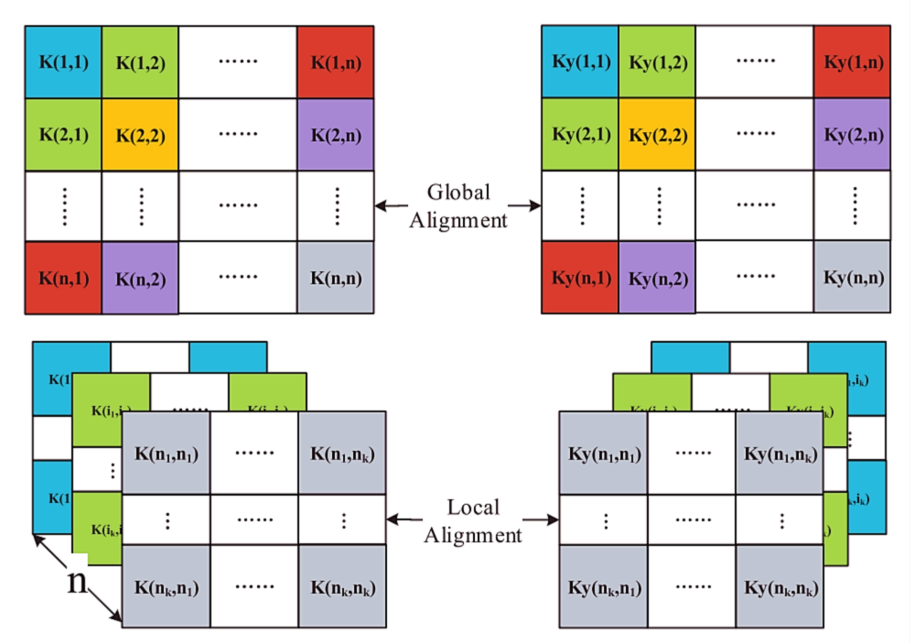
" Algorithms for Learning Kernels Based on Centered Alignment. " , JMLR 2012.

MKL (Classification Tasks)

- **MKL with Hybrid Kernel Alignment (HKTA)**

$$\boldsymbol{KTA}_{hybrid}(\boldsymbol{K}_1, \boldsymbol{K}_2) = (1-\lambda)\boldsymbol{KTA}_{local}(\boldsymbol{K}_1, \boldsymbol{K}_2) + \lambda\, \boldsymbol{KTA}_{global}(\boldsymbol{K}_1, \boldsymbol{K}_2)$$

$$\max_{\boldsymbol{w}\geq \boldsymbol{0}, \|\boldsymbol{w}\|_2=1} (1-\lambda)\Big(\frac{1}{n}\sum_{i=1}^{n} \frac{<\boldsymbol{K}_{\boldsymbol{w}}^{(i)}, \boldsymbol{K}_{\boldsymbol{y}}^{(i)}>_{\boldsymbol{F}}}{\left\|\boldsymbol{K}_{\boldsymbol{w}}^{(i)}\right\|_{\boldsymbol{F}}} + \lambda \frac{<\boldsymbol{K}_{\boldsymbol{w}}, \boldsymbol{K}_{\boldsymbol{y}}>_{\boldsymbol{F}}}{\|\boldsymbol{K}_{\boldsymbol{w}}\|_{\boldsymbol{F}}}$$



Local Kernel:

$$K^{(i)} \in R^{k\times k}$$

"Multiple kernel learning with hybrid kernel alignment maximization .", Pattern Recognition 2017.

Multiple Kernel K-Means (Clustering Tasks)

Kernel k-means (KKM)

$$\min_{\mathbf{Z}\in\{0,1\}^{n\times k}} \sum_{i=1,c=1}^{n,k} Z_{ic}\|\emptyset(\boldsymbol{x}_i)-\boldsymbol{\mu}_c\|_2^2 \quad s.t. \sum_{c=1}^{k} Z_{ic}=1\,,\boldsymbol{\mu}_c=\frac{1}{n_c}\sum_{i=1}^{n} Z_{ic}\emptyset(\boldsymbol{x}_i)$$

Reformulated equation

$$\min_{\boldsymbol{H}\in R^{n\times k}} Tr\left(\boldsymbol{K}(\boldsymbol{I}_n-\boldsymbol{H}\boldsymbol{H}^T)\right) \quad s.t.\ \boldsymbol{H}^T\boldsymbol{H}=\boldsymbol{I}_k$$

Multiple kernel k-means (MKKM)

$$K_w\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\emptyset_w(\boldsymbol{x}_i)^T\emptyset_w\left(\boldsymbol{x}_j\right)=\sum_{p=1}^{m} w_p^2 K_p\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)$$

$$\emptyset_w(\boldsymbol{x})=[w_1\emptyset_1(\boldsymbol{x})^T,w_2\emptyset_2(\boldsymbol{x})^T,\dots,w_m\emptyset_m(\boldsymbol{x})^T]^T$$

Multiple kernels

$$\min_{\boldsymbol{H}\in R^{n\times k},\boldsymbol{w}\in R^m} Tr\left(\boldsymbol{K}_w(\boldsymbol{I}_n-\boldsymbol{H}\boldsymbol{H}^T)\right)$$

$$s.t.\ \boldsymbol{H}^T\boldsymbol{H}=\boldsymbol{I}_k, \qquad \boldsymbol{w}^T\boldsymbol{1}_m=1,\boldsymbol{w}\geq\boldsymbol{0}$$

"Multiple Kernel k-Means Clustering with Matrix-Induced Regularization.“, AAAI 2016.

Kernel K-Means (Clustering Tasks)

Multiple kernel k-means with matrix-induced regularization (MKKM-MR)

$$\min_{\boldsymbol{H}\in R^{n\times k},\boldsymbol{w}\in R^{m}} Tr\left(\boldsymbol{K}_w(\boldsymbol{I}_n - \boldsymbol{H}\boldsymbol{H}^T)\right) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

$$s.t.\ \boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{I}_k, \boldsymbol{w}^T\boldsymbol{1}_m = 1, \boldsymbol{w} \geq 0$$

"Multiple Kernel k-Means Clustering with Matrix-Induced Regularization.", AAAI 2016.

Optimal neighborhood kernel clustering with multiple kernels

$$\min_{\boldsymbol{H}\in R^{n\times k},\boldsymbol{G}\in S_+^n,\boldsymbol{w}} Tr\left(\boldsymbol{G}(\boldsymbol{I}_n - \boldsymbol{H}\boldsymbol{H}^T)\right) + \frac{\rho}{2}\|\boldsymbol{G} - \boldsymbol{K}_w\|_F^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

$$s.t.\ \boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{I}_k, \boldsymbol{w}^T\boldsymbol{1}_m = 1, \boldsymbol{w} \geq \boldsymbol{0}, \boldsymbol{G} \succcurlyeq 0$$

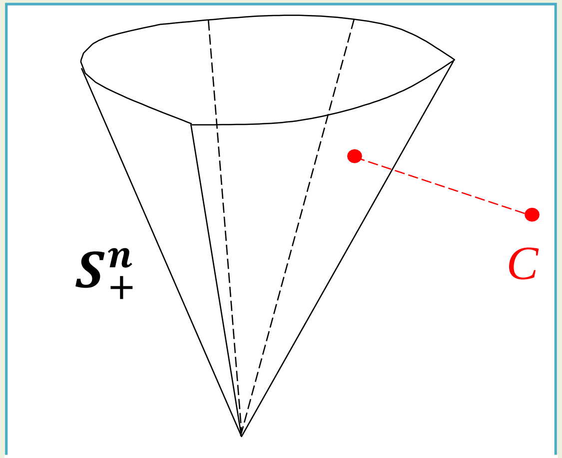
where

$$K_w(\boldsymbol{x}_i, \boldsymbol{x}_j) = \emptyset_w(\boldsymbol{x}_i)^T\emptyset_w(\boldsymbol{x}_j) = \sum_{p=1}^{m} w_p^2 K_p(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

"Optimal neighborhood kernel clustering with multiple kernels.", AAAI 2017.

$\boldsymbol{S}^n$: the space of symmetric matrices

$\boldsymbol{S}_+^n$: cone of PSD matrices



Projection onto the PSD Cone

References

- Schölkopf, B., Smola, A. J., & Bach, F. (**2018**). ”***Learning with kernels: support vector machines, regularization, optimization, and beyond***.” MIT press.
- Shawe-Taylor, J., & Cristianini, N. (**2004**). ”***Kernel methods for pattern analysis***.”, Cambridge university press.
- Moeller, J., Srikumar, V., Swaminathan, S., Venkatasubramanian, S., & Webb, D. (**2016**, September). “**Continuous kernel learning**.” In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 657-673). Springer, Cham.
- Huang, X., Suykens, J. A., Wang, S., Hornegger, J., & Maier, A. (**2017**). “**Classification With Truncated $\boldsymbol{l_1}$ Distance Kernel**.”, *IEEE transactions on neural networks and learning systems* (**TNNLS**), *29*(5), 2025-2030.
- Liu, F., Huang, X., Gong, C., Yang, J., & Suykens, J. A. (**2018**). “**Indefinite kernel logistic regression with concave-inexact-convex procedure.**”, *IEEE Transactions on Neural Networks and Learning Systems (**TNNLS**)*, *30*(3), 765-776.
- Gönen, M., & Alpaydın, E. (**2011**). **“Multiple kernel learning algorithms.”** *Journal of machine learning research (**JMLR**)*, *12*(Jul), 2211-2268.
- M. Gonen (**2012**), “**Bayesian efficient multiple kernel learning**”, ICML, 1206.6465.
- E.V. Strobl, S. Visweswaran, (**2013**) “**Deep multiple kernel learning**”, in: 12th International Conference on Machine Learning and Applications, vol. 1, IEEE, 2013, pp. 414–417.
- Aiolli, M. Donini, (**2015**) “**EasyMKL: a scalable multiple kernel learning algorithm**”, Neurocomputing 169 215–224.
- Y. Wang, X. Liu, Y. Dou, Q. Lv, Y. Lu, (**2017**) “**Multiple kernel learning with hybrid kernel alignment maximization**”, Pattern Recognit. 70, 104–111 .
- N. Cristianini, J. Kandola, A. Elisseeff, J. Shawe-Taylor, (**2006**) “**On kernel target alignment**”, in: Innovations in Machine Learning, Springer, pp. 205–256.
- M.P. Kumar, B. Packer, D. Koller, (**2010**), “**Self-paced learning for latent variable models**”, in: NIPS, vol. 1, p. 2.

References

- A.R. Meenakshi, C. Rajian, (**1999**) "**On a product of positive semidefinite matrices**", Linear Algebra Appl. 295 (1–3) 3–6.
- A.R. Conn, N.I.M. Gould, P.L. Toint, (**2000**) "**Trust Region Methods**", SIAM.
- Du, L., Zhou, P., Shi, L., Wang, H., Fan, M., Wang, W., Shen, Y.D., (**2015**) "**Robust multiple kernel k-means using l21-norm**", in: Twenty-fourth international joint conference on artificial intelligence, pp. 3476–3482.
- Gonen, M., Margolin, A.A., (**2014**) "**Localized data fusion for kernel k-means clustering with application to cancer biology**". Advances in Neural Information Processing Systems 27, 1305–1313.
- Huang, J., Nie, F., Huang, H., (**2015**) "**A new simplex sparse learning model to measure data similarity for clustering**", in: Twenty-fourth international joint conference on artificial intelligence.
- Yao, Y., Li, Y., Jiang, B., Chen, H., (**2020**) "**Multiple kernel k-means clustering by selecting representative kernels**". IEEE Transactions on Neural Networks and Learning Systems.
- Zhou, S., Liu, X., Li, M., Zhu, E., Liu, L., Zhang, C., Yin, J., (**2019**) "**Multiple kernel clustering with neighbor-kernel subspace segmentation**". IEEE transactions on neural networks and learning systems 31, 1351–1362.
- Shen, H.T., Zhu, Y., Zheng, W., Zhu, X., (**2020**) "**Half-quadratic minimization for unsupervised feature selection on incomplete data**". IEEE transactions on neural networks and learning systems 32, 3122–3135.
- Boyd, S., Parikh, N., Chu, E., (**2011**) "**Distributed optimization and statistical learning via the alternating direction method of multipliers**". Now Publishers Inc.
- Charbonnier, P., Blanc-F´eraud, L., Aubert, G., Barlaud, M., (**1997**). "**Deterministic edge-preserving regularization in computed imaging**". IEEE Transactions on image processing 6, 298–311.