Learning in Feature Spaces via Coupled Covariances: Asymmetric Kernel SVD and Nyström method

Qinghua Tao[1*], Francesco Tonin[2*], Alex Lambert[1], Yingyi Chen[1],Panagiotis Patrinos[1], and Johan A.K. Suykens[1]

[1]ESAT-STADIUS, KU Leuven, Belgium
[2]LIONS, EPFL, Switzerland (most work done at ESAT-STADIUS, KU Leuven)
[*]Equal contribution
qinghua.tao@esat.kuleuven.be, francesco.tonin@epfl.ch

ICML 2024

KU LEUVEN

European Research Council
Established by the European Commission

Motivation

Given $A \in \mathbb{R}^{n \times m}$, it can be seen as an array w.r.t. either rows or columns:

- $\mathcal{X} = \{A[i,:] \triangleq x_i\}_{i=1}^n$
- $\mathcal{Z} = \{A[:,j] \triangleq z_j\}_{j=1}^m$

SVD gives two sets of linear features for both $\mathcal{X}$ and $\mathcal{Z}$.

KPCA provides only one set of features to rows $\mathcal{X}$.

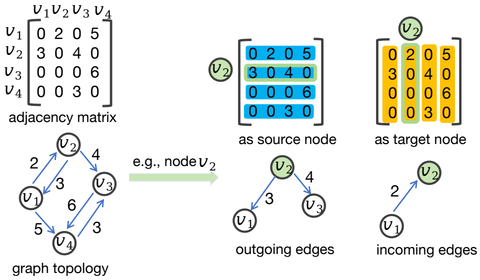


Figure: Example of asymmetric similarity.

- SVD can process any rectangular matrix, but lacks flexibility for nonlinearity.
- Classical kernel methods only deal with symmetric kernels.

Background: KSVD with LSSVMs Setups

Given two sets of samples $\{x_i \in \mathcal{X}\}_{i=1}^{n}, \{z_j \in \mathcal{Z}\}_{j=1}^{m}$ and feature mappings $\phi\colon \mathcal{X} \to \mathcal{H}, \psi\colon \mathcal{Z} \to \mathcal{H}$, the primal form of KSVD is given by

$$\begin{aligned} &\max_{w,v,e,r} -v^\top w + \frac{1}{2\lambda}\sum\nolimits_{i=1}^{n} e_i^2 + \frac{1}{2\lambda}\sum\nolimits_{j=1}^{m} r_j^2 \\ \text{s.t.} \quad & e_i = w^\top \phi(x_i),\ i = 1,\dots,n, \\ & r_j = v^\top \psi(z_j),\ j = 1,\dots,m, \end{aligned}$$

KSVD

The KKT conditions of KSVD leads to the shifted eigenvalue problem [1]:

$$G^\top B_\phi = B_\psi \Lambda, \quad G B_\psi = B_\phi \Lambda$$

where $G = [\frac{1}{\sqrt{nm}}\langle \phi(x_i), \psi(z_j)\rangle] \in \mathbb{R}^{n\times m}$ is an asymmetric kernel.

According to Lanczos' decomposition theorem [2], KSVD above can be solved by taking for B_ϕ, B_ψ the top-r left and right singular vectors of the matrix G.

[1] Suykens, J. A. SVD revisited: A new variational principle, compatible feature maps and nonlinear extensions. Applied and Computational Harmonic Analysis, 2016.

[2] Lanczos, C. Linear systems in self-adjoint form. The American Mathematical Monthly, 1958.

Main Goals

- **Coupled Covariance Eigenproblem (CCE)**
  - a new learning paradigm through covariance operators, complementing the kernel-based formulations for KSVD
  - allowing infinite-dimensional feature maps in KSVD
- **Asymmetric Nyström**
  - finite-sample approximation to integral equations w.r.t. asymmetric kernels and singular functions
  - faster computation for KSVD with large-scale kernels.

Coupled Covariance Eigenproblem (CCE)

In CCE, the goal is to **learn a pair of r directions in the feature space $\mathcal{H}$** solving a coupled eigenvalues problem. We define

- the sough-after directions in vectors of

$$W_\phi = [w_1^\phi, \ldots, w_r^\phi] \in \mathcal{H}^r, \quad W_\psi = [w_1^\psi, \ldots, w_r^\psi] \in \mathcal{H}^r,$$

- the empirical covariance operators

$$\Sigma_\phi = \frac{1}{n} \sum\nolimits_{i=1}^{n} \phi(x_i)\phi(x_i)^*, \quad \Sigma_\psi = \frac{1}{m} \sum\nolimits_{j=1}^{m} \psi(z_j)\psi(z_j)^*.$$

Definition (CCE)

Find $W_\phi \in \mathcal{H}^r$, $W_\psi \in \mathcal{H}^r$ such that

$$\Sigma_\phi W_\psi = \Lambda W_\phi, \qquad \Sigma_\psi W_\phi = \Lambda W_\psi,$$

for some diagonal matrix $\Lambda \in \mathbb{R}^{r \times r}$ with positive values.

Equivalence between CCE and KSVD

1. Given that a solution to the CCE exists, it holds that all directions $\{w_l^\phi\}_{l=1}^r$, $\{w_l^\psi\}_{l=1}^r$ lie respectively in Span $\{\phi(x_i)\}_{i=1}^n$, Span $\{\psi(z_j)\}_{j=1}^m$:

$$w_l^\phi = \sum_{i=1}^n b_{il}^\phi \phi(x_i), \qquad w_l^\psi = \sum_{j=1}^m b_{jl}^\psi \psi(z_j)$$

   where $B_\phi \in \mathbb{R}^{n\times r}$ and $B_\psi \in \mathbb{R}^{m\times r}$ denote the matrices of coefficients.

2. Let Γ_ϕ, Γ_ψ be linear operators on $W \in \mathcal{H}^r$ by $[\Gamma_\phi W]_{il} = \langle \phi(x_i), w_l \rangle / \sqrt{n}$, $[\Gamma_\psi W]_{jl} = \langle \psi(z_j), w_l \rangle / \sqrt{m}$, and $G = [\langle \phi(x_i), \psi(z_j) \rangle / \sqrt{nm}] \in \mathbb{R}^{n\times m}$. We have

$$W_\phi = \Gamma_\phi^* B_\phi, \qquad W_\psi = \Gamma_\psi^* B_\psi$$

$$\Gamma_\psi \Gamma_\phi^* B_\phi = G^\top B_\phi, \qquad \Gamma_\phi \Gamma_\psi^* B_\psi = G B_\psi$$

3. **Equivalence between the solutions in CCE and KSVD**

   Directions W_ϕ, W_ψ are solution to CCE if and only if B_ϕ, B_ψ are solution to:

$$G^\top G B_\psi = G^\top B_\phi \Lambda, \qquad G G^\top B_\phi = G B_\psi \Lambda,$$

Let B_ϕ^{svd} (resp. B_ψ^{svd}) be top-r left (resp. right) singular vectors of G from the KSVD. Then $W_\phi = \Gamma_\phi^* B_\phi^{svd}$, $W_\psi = \Gamma_\psi^* B_\psi^{svd}$ is a solution to the CCE.

Asymmetric Nyström Method

With an asymmetric kernel $\kappa(x, z)$, $u_s(x)$ and $v_s(z)$ satisfying

$$\lambda_s u_s(x) = \int_{\mathcal{D}_z} \kappa(x, z) v_s(z)\, p_z(z) dz,$$
$$\lambda_s v_s(z) = \int_{\mathcal{D}_x} \kappa(x, z) u_s(x)\, p_x(x) dx$$

are called a pair of **adjoint eigenfunctions** (singular functions) corresponding to the singular values λ_s with $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.

Through finite-sample approximation, the asymmetric Nyström gives:

$$\tilde{u}_s^{(N,M)} = (\sqrt{\sqrt{mn} l_{\lambda_s}} / \lambda_s^{(n,m)}) G_{N,m} v_s^{(n,m)},$$
$$\tilde{v}_s^{(N,M)} = (\sqrt{\sqrt{mn} l_{\lambda_s}} / \lambda_s^{(n,m)}) G_{n,M}^{\top} u_s^{(n,m)},$$

where $\lambda_s^{(n,m)}$, $u_s^{(n,m)}$, and $v_s^{(n,m)}$ are from the **SVD on an $n \times m$ (smaller) submatrix sampled from $G \in \mathbb{R}^{N \times M}$**.

[3] Williams, C. and Seeger, M. Using the Nyström method to speed up kernel machines. NeurIPS 2000.

Conclusion

- A new asymmetric learning paradigm with **CCE**, allowing infinite dimensional maps and providing covariance-based perspective for KSVD.
- Formal derivations to **asymmetric Nyström** method by starting from integral equations related to the continuous analogue of SVD.
- **Extensive experiments** on feature learning with asymmetric kernels.

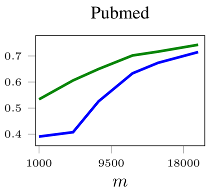
Dataset	F1 Score (↑)	PCA	KPCA	SVD	KSVD	DeepW	HOPE	DiGAE
Cora	Micro	0.757	0.771	0.776	**0.792**	0.741	0.750	0.783
	Macro	0.751	0.767	0.770	**0.784**	0.736	0.473	0.776
Citeseer	Micro	0.648	0.666	0.667	**0.678**	0.624	0.642	0.663
	Macro	0.611	0.635	0.632	**0.640**	0.587	0.607	0.627
Pubmed	Micro	0.765	0.754	0.766	0.773	0.759	0.771	**0.781**
	Macro	0.736	0.715	0.738	0.743	0.737	0.741	**0.749**
TwitchPT	Micro	0.681	0.681	0.694	**0.712**	0.637	0.685	0.633
	Macro	0.517	0.531	0.543	**0.596**	0.589	0.568	0.593
BlogCatalog	Micro	0.648	0.663	0.687	**0.710**	0.688	0.704	0.697
	Macro	0.643	0.659	0.673	**0.703**	0.679	0.697	0.690

- KSVD outperforms KPCA and even the methods specified for graphs

Method	ACM		DBLP		Pubmed		Wiki	
	NMI	Coh	NMI	Coh	NMI	Coh	NMI	Coh
SVD	0.58	0.21	0.09	-0.06	0.31	0.42	0.39	0.42
KPCA	0.59	0.28	0.26	0.17	0.29	0.51	0.46	0.57
KSVD	**0.68**	**0.32**	**0.28**	0.21	**0.33**	0.54	**0.48**	**0.64**
BCOT	0.38	0.27	0.27	**0.22**	0.16	0.54	**0.48**	**0.64**
EBC	0.62	0.20	0.15	0.21	0.19	**0.56**	0.47	0.63

- KSVD is comparable to the methods specified for bi-clustering

Task	N	Time (s) TSVD	RSVD	Sym. Nys.	Ours	Speedup
Cora	2708	0.841	0.274	0.673	**0.160**	1.71×
Citeseer	3312	0.568	0.290	0.214	**0.136**	2.14×
PubMed	19717	9.223	4.577	44.914	**0.141**	32.51×



- Asymmetric Nyström significantly speed up KSVD, and outperform symmetric Nyström with the same number of samplings.