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

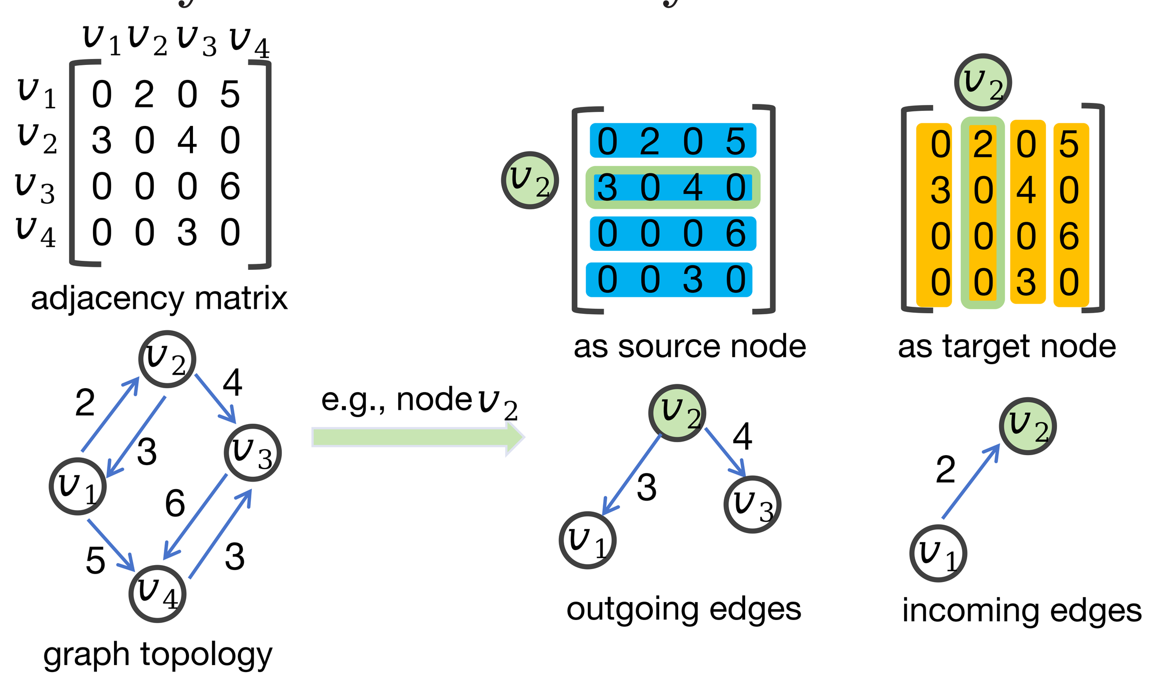
Qinghua Tao[1,*], Francesco Tonin[2,*], Alex Lambert[1], Yingyi Chen[1], Panagiotis Patrinos[1], Johan Suykens[1]

*Equal contribution [1]ESAT, KU Leuven, Belgium [2]LIONS, EPFL, Switzerland (most work done at ESAT-STADIUS, KU Leuven)

European Research Council
Established by the European Commission

Kernel SVD Problem

Example on Asymmetric Similarity:



Definition 1 (KSVD). Given two sets of samples $\{x_i \in \mathcal{X}\}_{i=1}^n, \{z_j \in \mathcal{Z}\}_{j=1}^m$ and feature mappings $\phi: \mathcal{X} \to \mathcal{H}, \psi: \mathcal{Z} \to \mathcal{H}$, the KKT conditions of KSVD under LSSVM setups leads to the shifted eigenvalue problem:

$$G^\top B_\phi = B_\psi \Lambda, \quad G B_\psi = B_\phi \Lambda \tag{1}$$

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

Theorem 2 (Decomposition Theorem, Lanczos). *Any nonzero matrix A can be written as $A = U\Sigma V^\top$, where U, V, Σ are defined by the shifted eigenvalue problem $A^\top U = V\Lambda, AV = U\Sigma$ [d].*

Current Limitations under LSSVM setups:

- only working with finite-dimensional feature mappings.
- possible unboundness in the variational objective.
- inefficiency with large-scale asymmetric kernels.

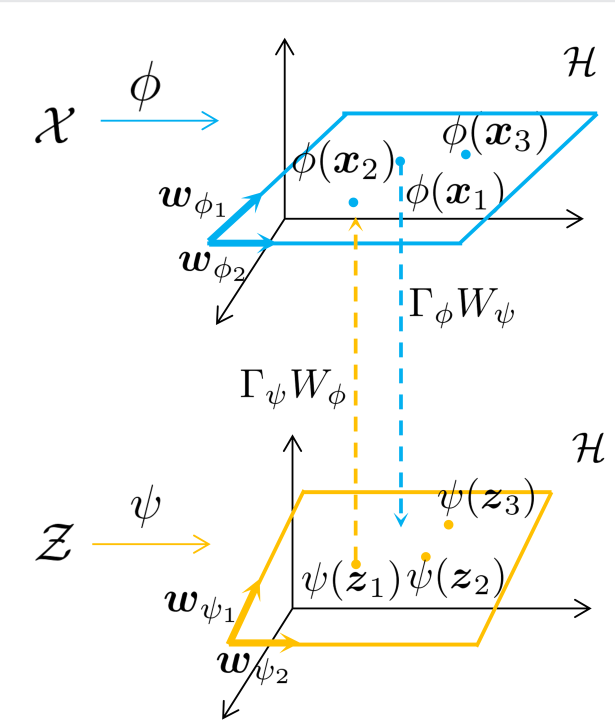
Paper TL;DR

We present CCE: a *coupled eigenvalue problem* that allows asymmetric learning in feature spaces, as well as a Nyström method for the corresponding asymmetric matrix.

Coupled Covariance EigenProblem

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
  $W_\phi = [w_1^\phi, \dots, w_r^\phi] \in \mathcal{H}^r$,
  $W_\psi = [w_1^\psi, \dots, w_r^\psi] \in \mathcal{H}^r$,
- the empirical covariance operators
  $\Sigma_\phi = \frac{1}{n} \sum_{i=1}^n \phi(x_i)\phi(x_i)^*$
  $\Sigma_\psi = \frac{1}{m} \sum_{j=1}^m \psi(z_j)\psi(z_j)^*$.



Definition 3 (CCE). Find $W_\phi \in \mathcal{H}^r, W_\psi \in \mathcal{H}^r$ such that

$$\Sigma_\phi W_\psi = \Lambda W_\phi, \qquad \Sigma_\psi W_\phi = \Lambda W_\psi, \tag{2}$$

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

CCE: KSVD via Covariance Operators

- 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 $\{\psi(z_j)\}_{j=1}^m$:

$$w_l^\phi = \sum_{i=1}^n b_{il}^\phi \phi(x_i), \qquad w_l^\psi = \sum_{j=1}^m b_{jl}^\psi \psi(z_j) \tag{3}$$

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

- Let Γ_ϕ and Γ_ψ be linear operators acting on $W \in \mathcal{H}^r$ by $[\Gamma_\phi W]_{il} = \frac{1}{\sqrt{n}} \langle \phi(x_i), w_l \rangle, [\Gamma_\psi W]_{jl} = \frac{1}{\sqrt{m}} \langle \psi(z_j), w_l \rangle$, we have:

$$W_\phi = \Gamma_\phi^* B_\phi, \qquad W_\psi = \Gamma_\psi^* B_\psi. \tag{4}$$

$$\Gamma_\psi \Gamma_\phi^* B_\phi = G^\top B_\phi, \qquad \Gamma_\phi \Gamma_\psi^* B_\psi = G B_\psi \tag{5}$$

$$G^\top G B_\psi = G^\top B_\phi \Lambda, \qquad G G^\top B_\phi = G B_\psi \Lambda, \tag{6}$$

- W_ϕ, W_ψ are solution to CCE if and only if B_ϕ, B_ψ are solution to (6).

Proposition 4. *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. (**Equivalence between CCE and KSVD.**)*

Difference with Symmetric Methods

- **With covariance:**

KPCA	KCCA	CCE
$\Sigma_\phi w_\phi = \lambda_\phi w_\phi$	$\Sigma_{\phi\psi} w_\psi = \lambda \Sigma_\phi w_\phi$	$\Sigma_\phi w_\psi = \lambda w_\phi$
$\Sigma_\psi w_\psi = \lambda_\psi w_\psi$	$\Sigma_{\psi\phi} w_\phi = \lambda \Sigma_\psi w_\psi$	$\Sigma_\psi w_\phi = \lambda w_\psi$

- **With kernel:**

KPCA	KCCA	KSVD
$K_\phi b_\phi = \lambda_\phi b_\phi$	$K_\psi b_\psi = \lambda (K_\phi + \rho_1 I) b_\phi$	$G b_\psi = \lambda b_\phi$
$K_\psi b_\psi = \lambda_\psi b_\psi$	$K_\phi b_\phi = \lambda (K_\psi + \rho_2 I) b_\psi$	$G^\top b_\phi = \lambda b_\psi$

where $\rho_1, \rho_2 > 0$ are regularization constants, $K_\phi = [\langle \phi(x_i), \phi(x_j) \rangle] \in \mathbb{R}^{n \times n}$ and $K_\psi = [\langle \psi(z_i), \psi(z_j) \rangle] \in \mathbb{R}^{m \times m}$, and KCCA requires $n = m$.

Asymmetric Nyström Approximation

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

$$\begin{aligned} \lambda_s u_s(x) &= \int_{\mathcal{D}_z} \kappa(x,z) v_s(z)\, p_z(z) dz, \\ \lambda_s v_s(z) &= \int_{\mathcal{D}_x} \kappa(x,z) u_s(x)\, p_x(x) dx \end{aligned} \tag{7}$$

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:

$$\begin{aligned} \tilde{u}_s^{(N,M)} &= (\sqrt{\sqrt{mn} l_{\lambda_s}} / \lambda_s^{(n,m)}) G_{N,m} v_s^{(n,m)}, \\ \tilde{v}_s^{(N,M)} &= (\sqrt{\sqrt{mn} l_{\lambda_s}} / \lambda_s^{(n,m)}) G_{n,M}^\top u_s^{(n,m)}, \end{aligned} \tag{8}$$

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 $G \in \mathbb{R}^{N \times M}$**.

Numerical Experiments

Node classification of Directed Graphs:

- KSVD outperforms KPCA and even the methods specified for graphs.

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

Bi-clustering:

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

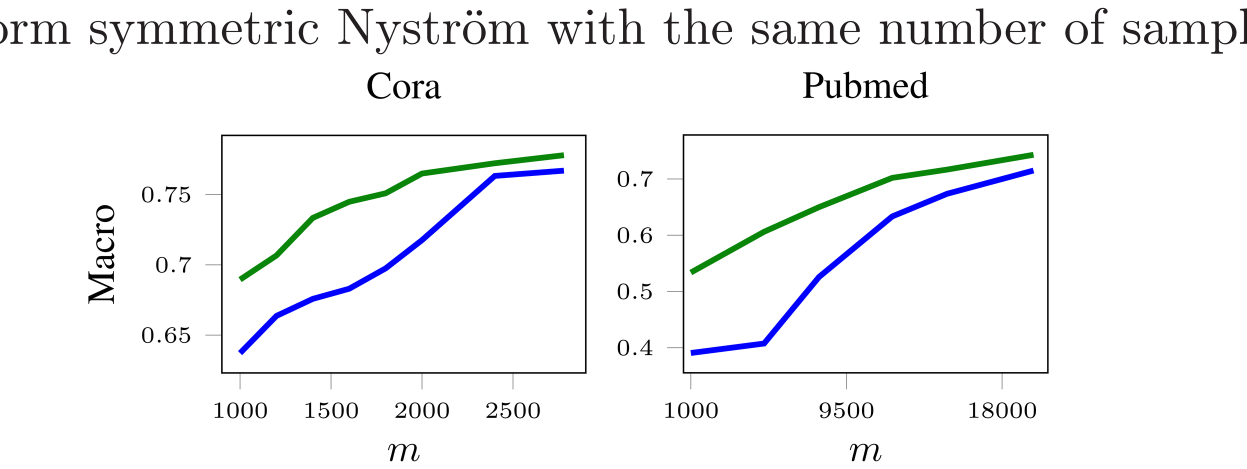
Method	ACM		DBLP		Pubmed		Wiki	
	NMI	Coh	NMI	Coh	NMI	Coh	NMI	Coh
SVD	0.58	0.21	0.09	-0.06	0.31	0.42	0.39	0.42
KPCA	0.59	0.28	0.26	0.17	0.29	0.51	0.46	0.57
KSVD	**0.68**	**0.32**	**0.28**	0.21	**0.33**	0.54	**0.48**	**0.64**
BCOT	0.38	0.27	0.27	**0.22**	0.16	0.54	**0.48**	**0.64**
EBC	0.62	0.20	0.15	0.21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

Asymmetric Nyström:

- Significantly speed up the computation of KSVD.

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×

- Outperform symmetric Nyström with the same number of samplings.



References

[a] Suykens, J. A. SVD revisited: A new variational principle, compatible feature maps and nonlinear extensions. ACHA, 2016.

[b] Chen, Y., Tao, Q., Tonin, F., and Suykens, J. A. Primal-attention: Self-attention through asymmetric kernel svd in primal representation. NeurIPS, 2023.

[c] Tao, Q., Tonin, F., Patrinos, P., and Suykens, J. A. Nonlinear SVD with Asymmetric Kernels: feature learning and asymmetric Nyström method. arXiv:2306.07040, 2023.

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

Contacts

qinghua.tao@esat.kuleuven.be, francesco.tonin@epfl.ch.