Extending Kernel PCA through Dualization: Sparsity, Robustness and Fast Algorithms

Francesco Tonin, Alex Lambert, Panagiotis Patrinos, Johan Suykens
ESAT, KU Leuven, Belgium

Kernel PCA Problem

Given: n datapoints $(x_i)_{i=1}^n \in \mathcal{X}^n$, feature map $\phi\colon \mathcal{X} \to \mathcal{H}$ to a Hilbert space $\mathcal{H}$.
Goal: find s directions in $\mathcal{H}$ that maximize the variance under orthonormal conditions. The KPCA optimization problem is

$$\sup_{W \in \mathcal{S}^s_{\mathcal{H}}} \frac{1}{2} \|\Gamma W\|^2_{\mathrm{F}}. \tag{1}$$

We use the following definitions.

- The Stiefel manifold of orthonormal s-frames in $\mathcal{H}$ is
$$\mathcal{S}^s_{\mathcal{H}} := \{W \in \mathcal{H}^s \mid \mathcal{G}(W) = I_s\}.$$
- $\mathcal{G}(W) \in \mathbb{R}^{s \times s}$ is the matrix such that $\mathcal{G}(W)_{ij} = \langle w_i, w_j \rangle$.
- $\Gamma\colon \mathcal{H}^s \to \mathbb{R}^{n \times s}$ is the linear operator s.t. for all $(i,j) \in [n \times s]$ and $W = (w_1, \dots, w_s) \in \mathcal{H}^s$, $[\Gamma W]_{ij} = \langle \phi(x_i), w_j \rangle$.
- G is the Gram matrix $G = [k(x_i, x_j)]_{i,j=1}^n$, where $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the positive definite kernel function induced by ϕ.

The usual way to solve (1) is through SVD of $G \Rightarrow$ slow with larger n.

Paper TL;DR

We propose a duality framework to solve the KPCA problem faster, with extension to robust and sparse losses.

Difference of convex functions

Key idea: Rewrite (1) as a difference of convex functions

$$\inf_{W \in \mathcal{H}^s} g(W) - f(\Gamma W), \tag{2}$$

with $f = \frac{1}{2}\|\cdot\|^2_{\mathrm{F}}$, $g = \iota_{\mathcal{S}^s_{\mathcal{H}}}(\cdot)$, and $\iota_{\mathcal{C}}(\cdot)$ the indicator function for set $\mathcal{C}$.

Two key advantages:

1. Allows new gradient-based algorithm to solve KPCA efficiently without the SVD of G.
2. It becomes possible to slightly modify the loss function f to enforce specific properties such as robustness or sparsity.

Proposition 0.1 (Dual of difference of convex functions)**.** *Let $\mathcal{U}, \mathcal{K}$ be two Hilbert spaces, $g\colon \mathcal{U} \to \bar{\mathbb{R}}$ and $f\colon \mathcal{K} \to \bar{\mathbb{R}}$ be two convex lower semi-continuous functions and $\Gamma \in \mathcal{L}(\mathcal{U}, \mathcal{K})$. The problem*

$$\inf_{W \in \mathcal{U}} g(W) - f(\Gamma W)$$

admits the dual formulation

$$\inf_{H \in \mathcal{K}} f^\star(H) - g^\star(\Gamma^\sharp H),$$

and strong duality holds.

Faster KPCA with Gradient Descent

Motivation for going from primal to dual: we show that $g^\star(\Gamma^\sharp H)$ is related to the nuclear norm of some low dimensional matrix.

Proposition 0.2. *Let g be the indicator function of the Stiefel manifold and Γ as in Problem 1. Then for all $H \in \mathbb{R}^{n \times s}$,*

$$g^\star(\Gamma^\sharp H) = \operatorname{Tr}\sqrt{H^\top G H} =: \pi(H).$$

The computational complexity of computing the gradient of π:

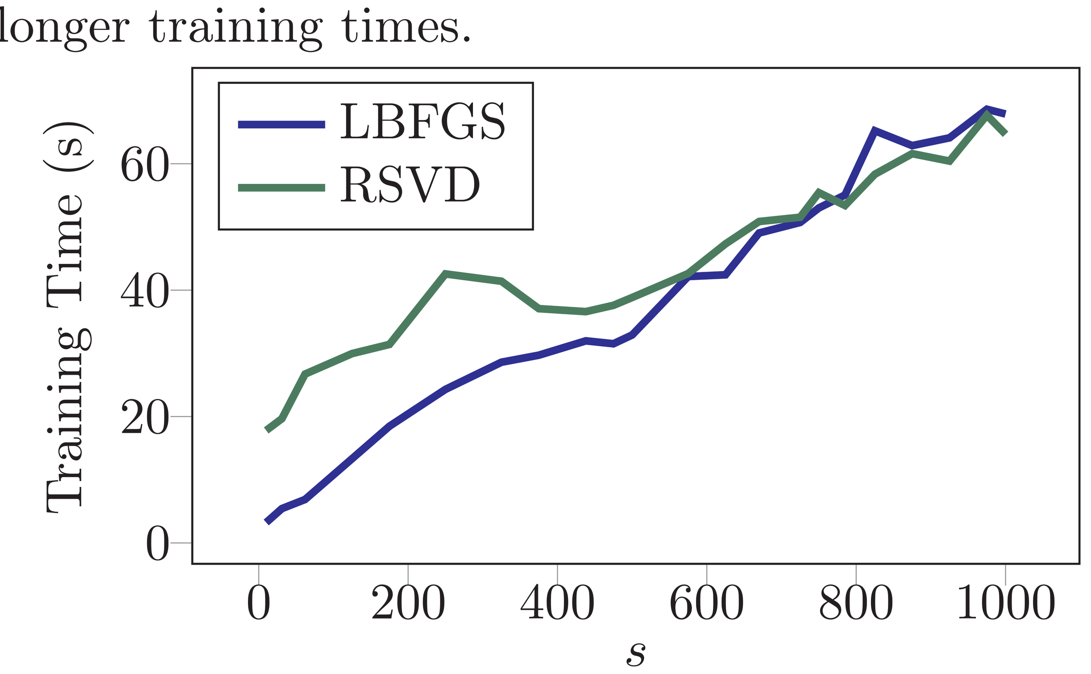
- Computation of $H^\top G H$ in $\mathcal{O}(sn^2)$,
- SVD of $H^\top G H$ in $\mathcal{O}(s^3)$.

We solve our dual problem with L-BFGS and compare training time with full SVD, Lanczos method, and Randomized SVD (RSVD).

- **KPCA Training Time** for multiple KPCA problems with fixed $\delta = 10^{-2}$ accuracy. Speedup factor w.r.t. RSVD.

Task	n	Time (s)				Speedup
		SVD	Lanczos	RSVD	Ours	Factor
Synth 1	7000	96.73	0.85	1.97	**0.53**	3.72
Protein	14895	868.64	3.46	6.70	**1.07**	6.25
RCV1	20242	-	6.04	12.50	**2.12**	5.90
CIFAR-10	60000	-	48.10	123.89	**13.51**	9.17

- **Influence of the number of components** s on training time: higher s leads to longer training times.



Beyond variance maximization

Typical loss function: square loss $f = \frac{1}{2}\|\cdot\|^2_{\mathrm{F}}$.
Problems: sensible to outliers, no sparsity.
Key idea: use a loss obtained with infimal convolution

$$f = \frac{1}{2}\|\cdot\|^2_{\mathrm{F}} \,\square\, \Psi,$$

where Ψ is a well-chosen function that enforces robustness or sparsity. Compatibility between the Fenchel-Legendre transform and the infimal convolution operator then allows to write the dual to Equation (2) as

$$\inf_{H \in \mathbb{R}^{n \times s}} \frac{1}{2}\|H\|^2_{\mathrm{F}} + \Psi^\star(H) - \pi(H).$$

DC Optimization

As f is a Moreau envelope, its gradient is always defined for all $Y \in \mathbb{R}^{n \times s}$,

$$\nabla\left(\frac{1}{2}\|\cdot\|^2_{\mathrm{F}} \,\square\, \Psi\right)(Y) = Y - \operatorname{prox}_{\Psi}(Y).$$

According to Moreau decomposition, it holds that for all $Y \in \mathbb{R}^{n \times s}$,

$$Y - \operatorname{prox}_{\Psi}(Y) = \operatorname{prox}_{\Psi^\star}(Y).$$

Algorithm 1 DCA for Moreau envelope objectives

input : Gram matrix G
for *epoch t from* 0 *to* $T-1$ **do**
  // alternating gradient steps
  $Y = \nabla\pi(H^{(t)})$
  $H^{(t+1)} = \operatorname{prox}_{\psi^\star}(Y)$
return $H^{(T)}$

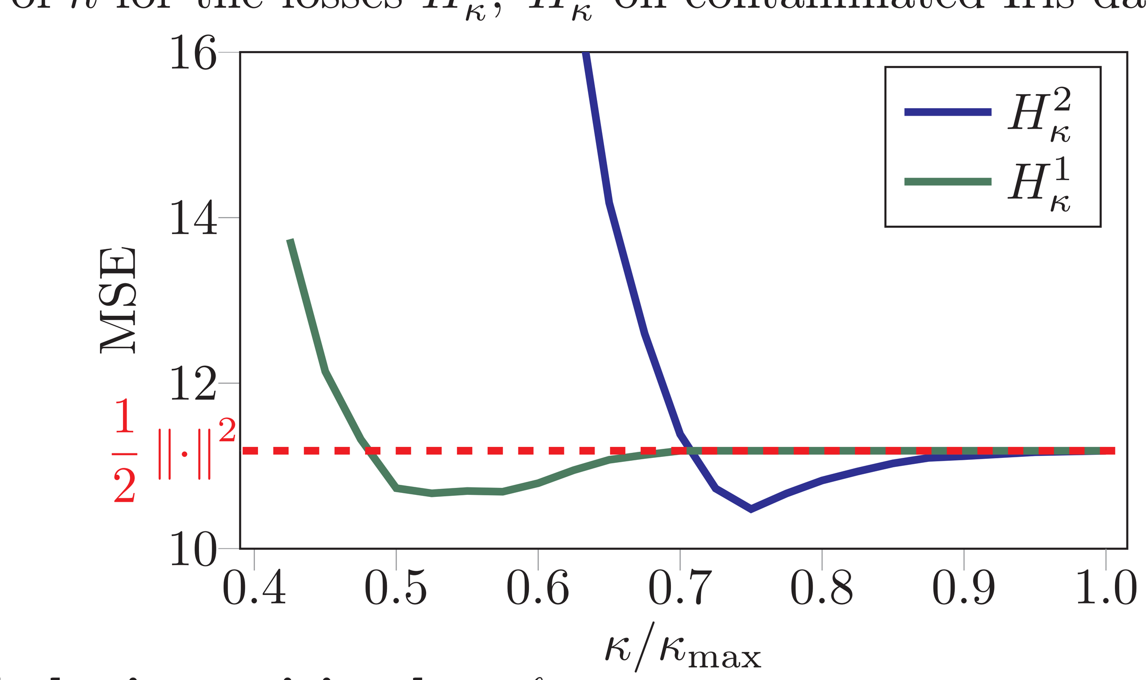
Robustness and sparsity

Denoting $\|\cdot\|_\star$ as the dual norm of $\|\cdot\|$ and the balls of radius t for these norms as $\mathcal{B}^\star_t$ and $\mathcal{B}_t$, we extend KPCA with Huber and ϵ-insensitive objectives to promote robustness and sparsity, respectively.

Extended Huber loss H_κ:

$$\Psi := \kappa\|\cdot\|, \qquad \Psi^\star = \iota_{\mathcal{B}^\star_\kappa}, \qquad \operatorname{prox}_{\Psi^\star}(Y) = \operatorname{Proj}_{\mathcal{B}^\star_\kappa}(Y).$$

- Effect of κ for the losses H^2_κ, H^1_κ on contaminated Iris dataset.



Extended ϵ-insensitive loss ℓ_ϵ:

$$\Psi := \iota_{\mathcal{B}_\epsilon}, \qquad \Psi^\star = \epsilon\|\cdot\|_\star, \qquad \operatorname{prox}_{\Psi^\star}(Y) = Y - \operatorname{Proj}_{\mathcal{B}_\epsilon}(Y).$$

- Reconstruction error for the ℓ^∞_ϵ loss for multiple ϵ and s.