G. V. Kriukova

On the Robustness of Kernel Principal Component Analysis

🏛 National University of “Kyiv-Mohyla Academy”, Kyiv.

Principal component analysis (PCA) is one of the primary statistical techniques for unsupervised linear dimension reduction and feature extraction. Kernel principal component analysis (KPCA) may be considered as a generalization of linear PCA that allows non-linear feature extraction [1]. The kernel represents an implicit mapping of the data space to a usually higher dimensional space, where linear PCA is performed. It consists of the eigenvalue decomposition calculation or singular value decomposition of centered kernel data, and quantification of orthogonal functions that optimize the kernel data scatter.

Namely, let $\mathcal{X}$ denote the data space, and $\mathcal{H}$ the feature space. The function $\varphi : \mathcal{X} \to \mathcal{H}$ is induced by a Mercer kernel $k : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$. It is known [1], that if $k(\cdot,\cdot)$ is a kernel, then the function $\varphi(\cdot)$ and the feature space $\mathcal{H}$ exist, moreover, $k(x,y) = \langle\varphi(x),\varphi(y)\rangle$. Let $\mathbf{K}$ denote the kernel matrix, i.e. $\mathbf{K}_{ij} = k(x_i, x_j)$ for items x_i, x_j from training set. KPCA is computed by finding first m eigenvectors corresponding to the largest eigenvalues of the kernel matrix $\mathbf{K}$. The eigenvectors in the feature space V can be computed as $V = \Gamma A$, where $\Gamma = [\varphi(x_1),\ldots,\varphi(x_n)]$, and A matrix of eigenvectors.

It is well known that the classical PCA and KPCA are not robust against data corruption, and even a small number of outliers can disturb the resulting principal components. Attempts to overcome this issue led to the development of various robustified procedures [2; 3]. However, due to the implicitness of the feature space, some extensions of PCA such as robust PCA cannot be directly generalized to KPCA.

We discuss an approach to treating noise, missing data, and outliers in KPCA, the pros and cons of various robustified KPCA algorithms, compare computational complexity and performance.

[1] B. Schölkopf, A. Smola, K.–R. Müller, *Nonlinear Component Analysis as a Kernel Eigenvalue Problem*, Neural Computation, vol. 10, no. 5, pp. 1299–1319, MIT, 1998. doi:10.1162/089976698300017467

[2] H. Kamiya, S. Eguchi, *A class of robust principal component vectors*, Journal of Multivariate Analysis, vol. 2, no. 77, pp. 239–269, 2001.

[3] M. H. Nguyen, F. Torre, *Robust Kernel Principal Component Analysis*, Advances in Neural Information Processing Systems, vol. 21, 2008.

E-mail: ✉ *kriukovagv@ukma.kiev.ua.*