KU LEUVEN

erc European Research Council

Henri De Plaen and Johan A. K. Suykens
ESAT-STADIUS, KU Leuven, Belgium

A Dual Formulation for Probabilistic Principal Component Analysis

Context & Contributions

- We characterize Probabilistic Principal Component Analysis (Tipping & Bishop, 1999) in Hilbert spaces and demonstrate how the **optimal solution admits a representation in dual space**.
- We develop a new extension of **KPCA** (Mika et al., 1998; Schölkopf et al., 1998) incorporating a **noise assumption** on the feature map.
- We give a **probabilistic** interpretation to the **generation in KPCA** (Schreurs & Suykens, 2018).
- We illustrate how the dual model works on a toy and a real dataset and **show its connections to KPCA**.

Definitions

As some kernels lead to **infinite dimensional feature maps**, we need to carefully define **finite subspaces** to allow the proper definition of **probability distributions**. Given a set of observations $\{\varphi_i \in \mathcal{H}\}_{i=1}^N$ and a kernel space $\mathcal{E}$ with basis $\{e_i\}_{i=1}^N$ and a finite feature space $\mathcal{H}_\mathcal{E} = \mathrm{span}\{\varphi_1, \dots, \varphi_N\}$. We define the mapping $\mathbf{\Phi} : \mathcal{E} \to \mathcal{H}_\mathcal{E} : \sum_{i=1}^N \varphi_i e_i^*$. This defines the covariance $\mathbf{\Phi} \circ \mathbf{\Phi}^*$ and the kernels $\mathbf{\Phi}^* \circ \mathbf{\Phi}$. We consider a latent space $\mathcal{L} \subset \mathcal{E}$ of dimension q and define the interconnection operator $\boldsymbol{W} : \mathcal{L} \to \mathcal{H}_\mathcal{L}$, where $\mathcal{H}_\mathcal{L} \subset \mathcal{H}_\mathcal{E}$. We refer to the feature spaces $\mathcal{H}$, $\mathcal{H}_\mathcal{E}$ and $\mathcal{H}_\mathcal{L}$ as the **primal spaces** and $\mathcal{E}$ and $\mathcal{L}$ as the **dual spaces**.

Distribution	Interpretation	Primal (features)	Dual (kernels)
latent | observation	latent projection	$\boldsymbol{h}\|\boldsymbol{\phi} \sim \mathcal{N}\left(\boldsymbol{\Sigma}_{\boldsymbol{h}\|\boldsymbol{\phi}}^{-1} \circ \boldsymbol{W}^*\boldsymbol{\phi}, \sigma^2\boldsymbol{\Sigma}_{\boldsymbol{h}\|\boldsymbol{\phi}}^{-1}\right)$	$\boldsymbol{h}\|\boldsymbol{k} \sim \mathcal{N}\left(\boldsymbol{\Sigma}_{\boldsymbol{h}\|\boldsymbol{k}}^{-1} \circ \boldsymbol{A}\boldsymbol{k}, \boldsymbol{\Sigma}_{\boldsymbol{h}\|\boldsymbol{k}}^{-1}\right)$
observation | latent	latent-based generation	$\boldsymbol{\phi}\|\boldsymbol{h} \sim \mathcal{N}\left(\boldsymbol{W}\boldsymbol{h}, \sigma^2\boldsymbol{I}_{\mathcal{H}_\mathcal{E}}\right)$	$\boldsymbol{k}\|\boldsymbol{h} \sim \mathcal{N}\left((\mathbf{\Phi}^* \circ \mathbf{\Phi}) \circ \boldsymbol{A}\boldsymbol{h}, \sigma^2\mathbf{\Phi}^* \circ \mathbf{\Phi}\right)$
latent	latent prior	$\boldsymbol{h} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_\mathcal{L})$	$\boldsymbol{h} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_\mathcal{L})$
observation	absolute generation	$\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{W} \circ \boldsymbol{W}^* + \sigma^2\boldsymbol{I}_{\mathcal{H}_\mathcal{E}})$	$\boldsymbol{k} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{A}^* \circ \boldsymbol{A} + \sigma^2(\mathbf{\Phi}^* \circ \mathbf{\Phi})^{-1}\right)$

Table: Interpretation of the different distributions of the Prob. PCA framework after training, in both primal and dual formulations. The covariance operators are given by $\boldsymbol{\Sigma}_{\boldsymbol{h}|\boldsymbol{\phi}} = (\boldsymbol{W}^* \circ \boldsymbol{W} + \sigma^2\boldsymbol{I}_\mathcal{L})^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{h}|\boldsymbol{k}} = \left(\boldsymbol{A}^* \circ (\mathbf{\Phi}^* \circ \mathbf{\Phi}) \circ \boldsymbol{A} + \sigma^2\boldsymbol{I}_\mathcal{L}\right)^{-1}$. *For the simplicity of this presentation, we do not consider the centering of kernels of feature maps and refer to the paper for these considerations.*

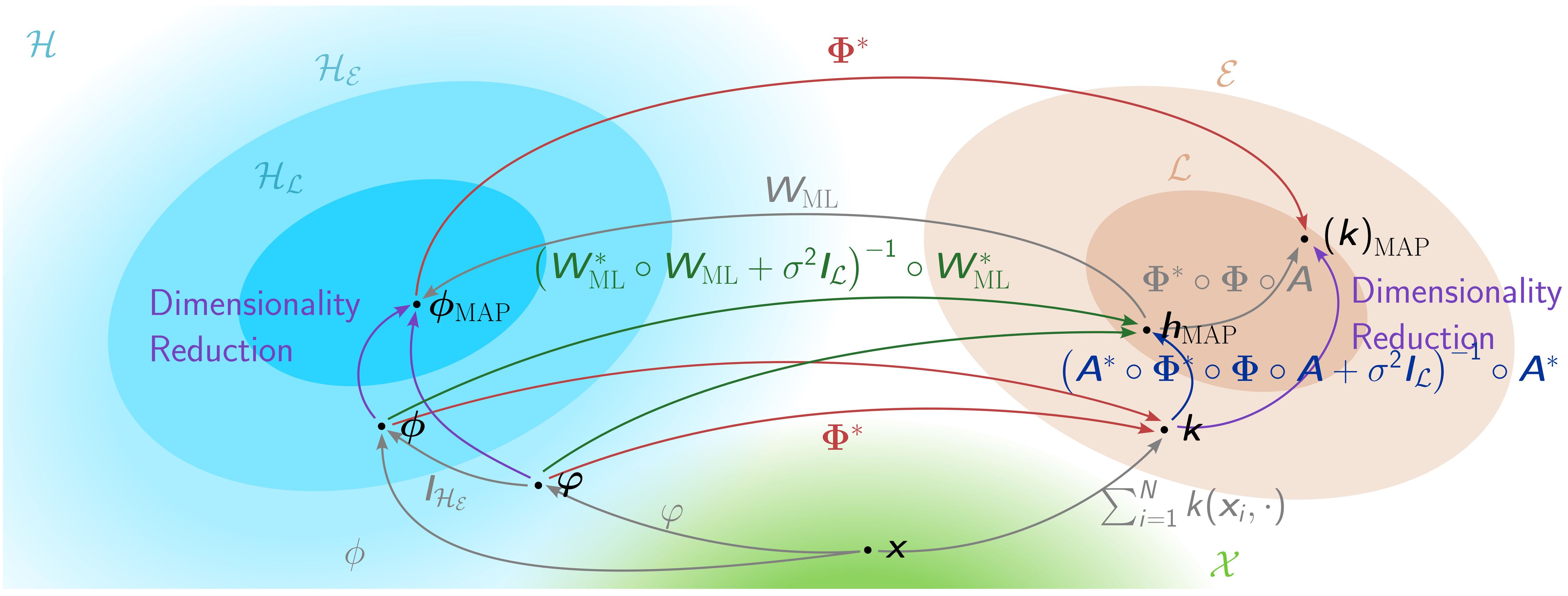


Figure: Global overview of the Probabilistic Principal Component Analysis in both primal and dual formulations. The primal spaces, or feature $\mathcal{H}$, $\mathcal{H}_\mathcal{E}$ and $\mathcal{H}_\mathcal{L}$ are in blue. The dual, or kernel and latent spaces $\mathcal{E}$ and $\mathcal{L}$ are in brown. The input space $\mathcal{X}$ is in green. The color or the applications (arrows) is just for the readability and has nothing to do with the color of the spaces.

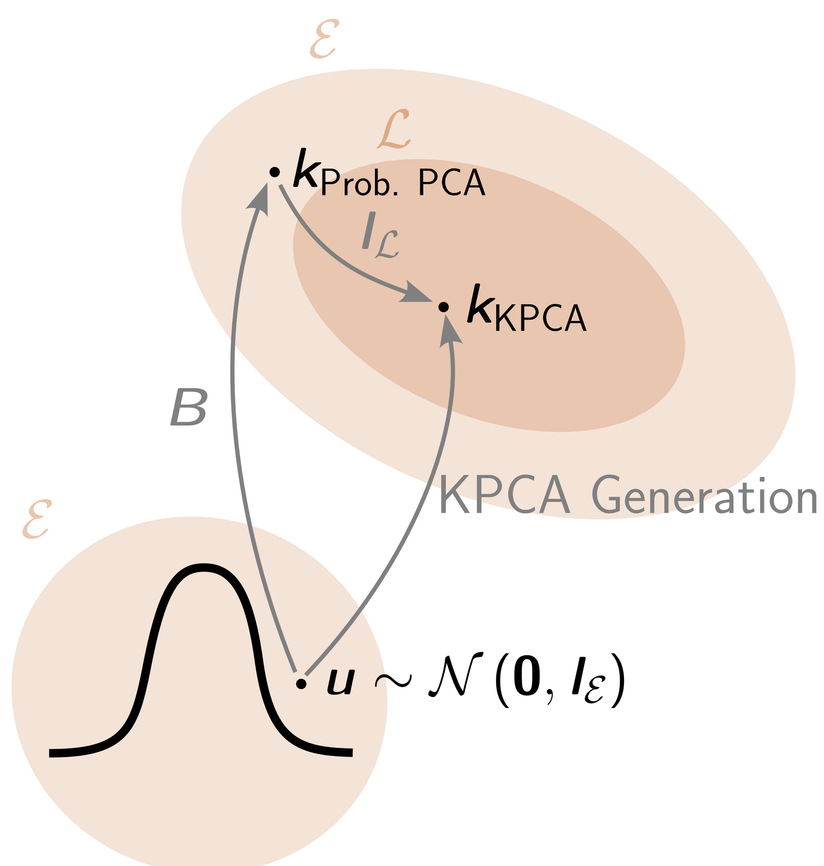


Figure: Schematic overview of the dual sampling in Prob. PCA compared to the generation in KPCA, with $B : \mathcal{E} \to \mathcal{E} : N^{-1/2}\sum_{p=1}^q \lambda_p \boldsymbol{\epsilon}_p \boldsymbol{r}_p^* + \sum_{p=q+1}^N \sigma\lambda_p^{1/2}\boldsymbol{\epsilon}_p\boldsymbol{r}_p^*$.

Maximum Likelihood

If we consider the eigendecomposition of the covariance $\mathbf{\Phi} \circ \mathbf{\Phi}^* = \sum_{i=1}^N \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^*$ and take the q dominant eigenpairs ($\lambda_1 \geq \dots \geq \lambda_q \geq \dots \geq \lambda_N$), we have

$$\boldsymbol{W}_{\mathrm{ML}} = \sum_{p=1}^q \sqrt{\lambda_p/N - \sigma_{\mathrm{ML}}^2}\,\boldsymbol{v}_p\boldsymbol{r}_p^*,$$
$$\sigma_{\mathrm{ML}}^2 = \frac{1}{N(N-q)}\sum_{p=q+1}^N \lambda_p,$$

with $\{\boldsymbol{r}_p\}_{p=1}^q$ and arbitrary orthonormal base of the latent space $\mathcal{L}$. We note that $\boldsymbol{W}_{\mathrm{ML}}$ is not unique as it is rotational invariant.

Representation

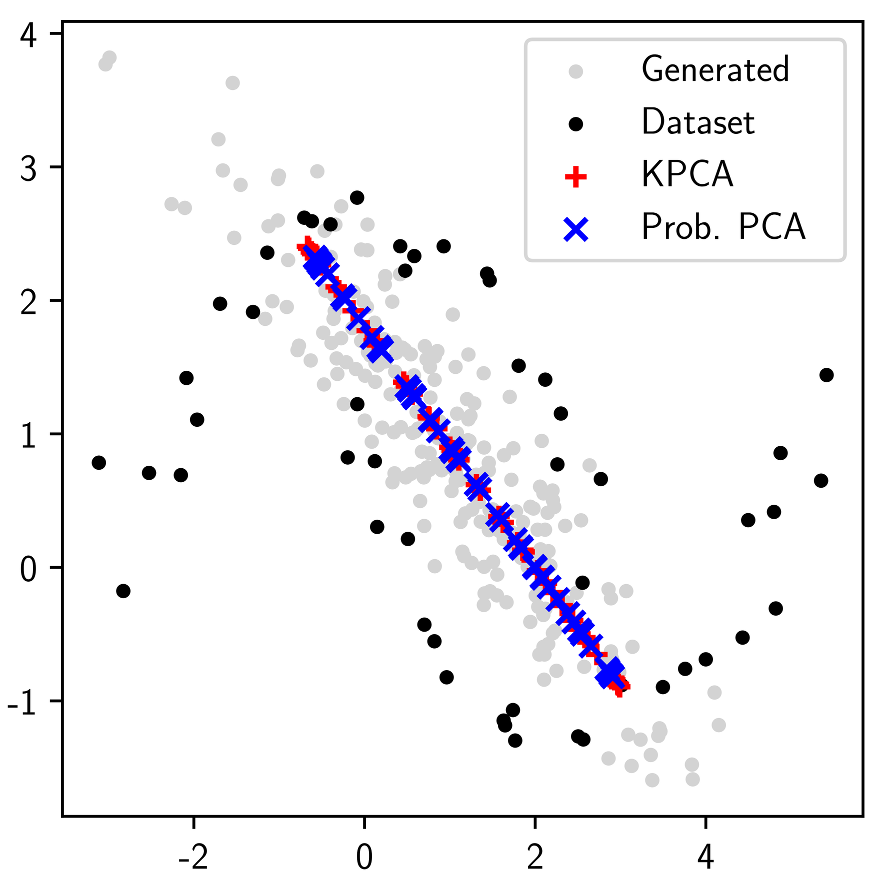
As a consequence of the choice of $\mathcal{H}_\mathcal{E}$ as $\mathrm{span}\{\varphi_1, \dots, \varphi_N\}$, we have

$$\boldsymbol{W} = \mathbf{\Phi} \circ \boldsymbol{A}, \qquad \text{with } \boldsymbol{A} : \mathcal{L} \to \mathcal{L},$$

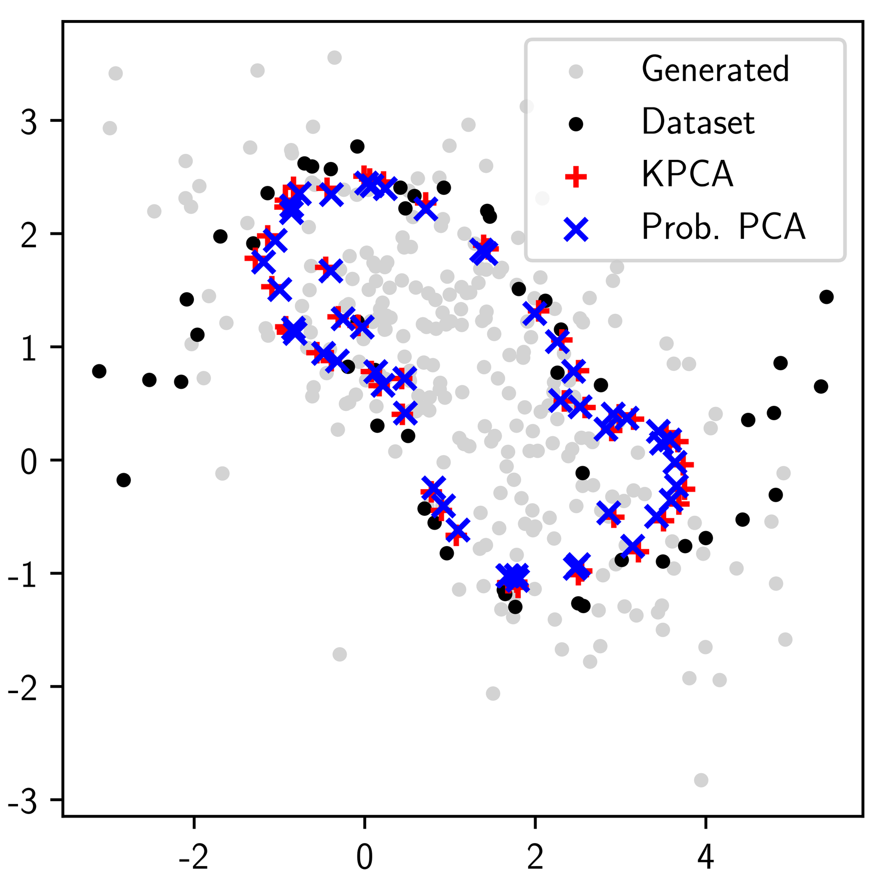
and in particular, as $\mathbf{\Phi}^* \circ \mathbf{\Phi}$ and $\mathbf{\Phi} \circ \mathbf{\Phi}^*$ share the same spectrum:

$$\boldsymbol{A}_{\mathrm{ML}} = \sum_{p=1}^q \sqrt{1/N - \sigma_{\mathrm{ML}}^2\lambda_p^{-1}}\,\boldsymbol{\epsilon}_p\boldsymbol{r}_p^*,$$

with $\{(\lambda_p, \boldsymbol{\epsilon}_p)\}_{p=1}^q$ the q dominant eigenpairs of $\mathbf{\Phi}^* \circ \mathbf{\Phi}$.



(a) With $q = 1$ component, the explained variance is 31.23% and $\sigma_{\mathrm{ML}}^2 = 1.40\%$.



(b) With $q = 3$ components, the explained variance is 54.03% and $\sigma_{\mathrm{ML}}^2 = 0.98\%$.

Figure: Visualisation of the Probabilistic PCA reconstruction (in blue) the classical KPCA (in red). Samples generated by are also given (in grey). The dataset contains $N = 20$ points (in black).

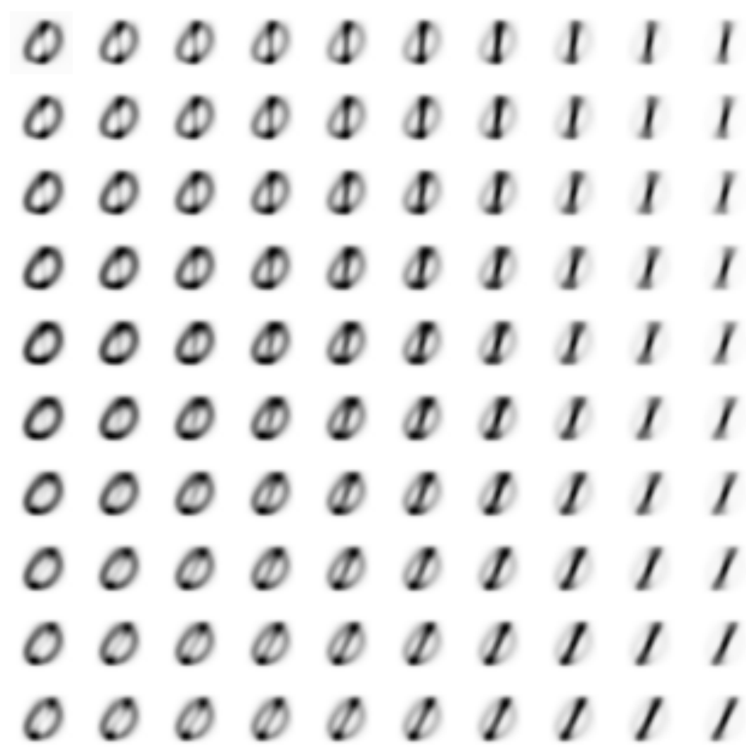


Figure: Generated datapoints on the MNIST dataset restricted to 0's and 1's, with $N = 500$ datapoints, with $q = 2$ components. The sample $\boldsymbol{u}$ is uniform on $[-1, 1]$ for the two first components and zero for the others. The horizontal axis varies in the first component and the vertical one in the second component. The explained variance is 27.97% and $\sigma_{\mathrm{ML}}^2 = 0.14\%$.