OLS Regression and PCA Result

Problem Setup: Let $X \in \mathcal{R}^{n \times d}$ be the features, $y \in \mathcal{R}^{n \times 1}$ be the labels. Assume X is full rank, implying that $\text{rank}(X) = \text{rank}(X^T X) = d$ and $n \geq d$.

Let the eigendecomposition of the covariance matrix be $X^T X = P^T \Lambda P$ i.e. the rows of $P \in \mathcal{R}^{d \times d}$ are the eigenvectors of the covariance matrix. Without loss of generality, assume that the rows are ordered with respect to the magnitude of the eigenvectors (the first row is the eigenvector corresponding to the largest eigenvalue, etc.) and that the eigenvectors are normalized, implying that $PP^T = I$.

Let $P_{k \times d}$ be the top-k rows of P with $k < d$. To perform Principal Component Analysis (PCA) on the data, compute $X P_{k \times d}^T$.

The ordinary-least squares (OLS) linear regression weights are given by:

$$w = (X^T X)^{-1} X^T y$$

Statement to Prove: We are concerned with Principal Component Regression (PCR) which involves reducing the dimensionality of a dataset using PCA, then learning an OLS model on the reduced data. We will show that the following two procedures result in the same learned weights:

1. (LHS) Learning weights in the original feature space, then applying PCA to the learned weights
2. (RHS) Learning OLS weights in the PCA-reduced feature space

Proof: We wish to show that $w_1 = w_2$ where

$$w_2 = P_{k \times d}(X^T X)^{-1} X^T y \qquad \text{and} \qquad w_1 = ((X P_{k \times d}^T)^T (X P_{k \times d}^T))^{-1} (X P_{k \times d}^T)^T y$$

Simplifying w_2:

$$\begin{aligned}
w_2 &= P_{k \times d}(X^T X)^{-1} X^T y && (1)\\
&= P_{k \times d}(P_{d \times d}^T \Lambda_{d \times d} P_{d \times d})^{-1} X^T y && (2)\\
&= P_{k \times d} P_{d \times d}^T \Lambda_{d \times d}^{-1} P_{d \times d} X^T y && (3)\\
&= I_{k \times d} \Lambda_{d \times d}^{-1} P_{d \times d} X^T y && (4)\\
&= \Lambda_{k \times d}^{-1} P_{d \times d} X^T y && (5)\\
&= \Lambda_{k \times k}^{-1} P_{k \times d} X^T y && (6)
\end{aligned}$$

Simplifying w_1:

$$
\begin{aligned}
w_1 &= ((XP_{k\times d}^T)^T(XP_{k\times d}^T))^{-1}(XP_{k\times d}^T)^T y && (7)\\
&= (P_{k\times d}X^TXP_{k\times d}^T)^{-1}P_{k\times d}X^Ty && (8)\\
&= (P_{k\times d}P_{d\times d}^T\Lambda_{d\times d}P_{d\times d}P_{k\times d}^T)^{-1}P_{k\times d}X^Ty && (9)\\
&= (I_{k\times d}\Lambda_{d\times d}I_{d\times k})^{-1}P_{k\times d}X^Ty && (10)\\
&= \Lambda_{k\times k}^{-1}P_{k\times d}X^Ty && (11)
\end{aligned}
$$

Therefore $w_1 = w_2 = \Lambda_{k\times k}^{-1}P_{k\times d}X^Ty$, and both procedures result in the same learned weights.