Let's start with SVD decomposition of $X = UDV^T$.

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ x_{31} & x_{32} & \dots & x_{3k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ u_{31} & u_{32} & \dots & u_{3n} \\ \vdots & \vdots & & \vdots \\ u_{n1} & u_{n2} & \dots & u_{nn} \end{pmatrix} \cdot \begin{pmatrix} d_{11} & & & \\ & d_{22} & & \\ & & \ddots & \\ & & & d_{kk} \\ & & & \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1k} \\ v_{21} & v_{22} & \dots & v_{2k} \\ \vdots & \vdots & & \vdots \\ v_{k1} & v_{k2} & \dots & v_{kk} \end{pmatrix}^T$$

Blank entries are zeros! The singular values on the diagonal of D are positive and are sorted from largest to lowest, $d_{11} > d_{22} > \dots > d_{kk}$.

The columns of the matrix $P = UD = XV$ are called principal components, $p_j = d_{jj}u_j$.

Let $\hat{D}$ be the modified D matrix, where we keep only the first r positive elements on the diagonal, d_{11}, ..., d_{kk} and replace the other elements by 0.

For $r = 2$ the approximation of the original X matrix may be written as $\hat{X} = U\hat{D}V^T$,

$$\begin{pmatrix} \hat{x}_{11} & \hat{x}_{12} & \dots & \hat{x}_{1k} \\ \hat{x}_{21} & \hat{x}_{22} & \dots & \hat{x}_{2k} \\ \hat{x}_{31} & \hat{x}_{32} & \dots & \hat{x}_{3k} \\ \vdots & \vdots & & \vdots \\ \hat{x}_{n1} & \hat{x}_{n2} & \dots & \hat{x}_{nk} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ u_{31} & u_{32} & \dots & u_{3n} \\ \vdots & \vdots & & \vdots \\ u_{n1} & u_{n2} & \dots & u_{nn} \end{pmatrix} \cdot \begin{pmatrix} d_{11} & & & \\ & d_{22} & & \\ & & & \\ & & & \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1k} \\ v_{21} & v_{22} & \dots & v_{2k} \\ \vdots & \vdots & & \vdots \\ v_{k1} & v_{k2} & \dots & v_{kk} \end{pmatrix}^T$$

The approximation $\hat{X} = U\hat{D}V^T$ may be also written as $\hat{X} = U_*D_*V_*^T$,

$$\begin{pmatrix} \hat{x}_{11} & \hat{x}_{12} & \dots & \hat{x}_{1k} \\ \hat{x}_{21} & \hat{x}_{22} & \dots & \hat{x}_{2k} \\ \hat{x}_{31} & \hat{x}_{32} & \dots & \hat{x}_{3k} \\ \vdots & \vdots & & \vdots \\ \hat{x}_{n1} & \hat{x}_{n2} & \dots & \hat{x}_{nk} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \\ \vdots & \vdots \\ u_{n1} & u_{n2} \end{pmatrix} \cdot \begin{pmatrix} d_{11} & \\ & d_{22} \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \\ \vdots & \vdots \\ v_{k1} & v_{k2} \end{pmatrix}^T$$

The matrix U_* consists of the leftmost r columns of U. The matrix D_* consists of the top-left $r \times r$ corner of D. The matrix V_*^T consists of the top r rows of V^T.

Errors of approximation are stored in the matrix:

$$X - \hat{X} = UDV^T - U\hat{D}V^T = U(D - \hat{D})V^T$$

Let's calculate the quality of this approximation.

We'll use two facts from linear algebra: $\text{trace}(LR) = \text{trace}(RL)$ and $\text{trace}(XX^T) = \text{trace}(X^TX) = \sum x_{ij}^2$.

$$\sum_{ij} x_{ij}^2 = \text{trace}(X^TX) = \text{trace}(VD^TU^TUDV^T) = \text{trace}(DD^TU^TU) = \text{trace}(D^TD) = \sum_i d_{ii}^2$$

By the same argument

$$\sum_{ij} p_{ij}^2 = \text{trace}(P^TP) = \text{trace}(D^TU^TUD) = \text{trace}(D^TD) = \sum_i d_{ii}^2$$

Recall that all vectors in the X matrix are standardized. Sample variance of a typical vector x_j is given by $\sum_i (x_i j - 0)^2/(n-1) = 1$. Hence $\|x_j\|^2 = \sum_i x_{ij}^2 = (n-1)$.

Hence,

$$\sum_{ij} x_{ij}^2 = \|x_1\|^2 + \|x_2\|^2 + \cdots + \|x_k\|^2 = k(n-1) = \sum_{i=1}^{k} d_{ii}^2;$$

Or,

$$\sum_{ij} p_{ij}^2 = \|p_1\|^2 + \|p_2\|^2 + \cdots + \|p_k\|^2 = k(n-1) = \sum_{i=1}^{k} d_{ii}^2;$$

The sum of squares of all approximation errors:

$$Q = \sum_{ij} (x_{ij} - \hat{x}_{ij})^2 = \text{trace}((X - \hat{X})(X - \hat{X})^T)$$

What is inside the trace?

$$(X - \hat{X})(X - \hat{X})^T = V(D - \hat{D})^T U^T U (D - \hat{D}) V^T = V(D - \hat{D})^T (D - \hat{D}) V^T$$

Let's finalize the calculation of total approximation error Q:

$$Q = \text{trace}((X - \hat{X})(X - \hat{X})^T) = \text{trace}(V(D - \hat{D})^T (D - \hat{D}) V^T) = \text{trace}((D - \hat{D})^T (D - \hat{D}) V^T V) =$$
$$= \text{trace}((D - \hat{D})^T (D - \hat{D})) = \sum_{i=r+1}^{k} d_{ii}^2 = k(n-1) - \sum_{i=1}^{r} d_{ii}^2;$$

Let's regress all the columns of X onto predictors U_* and obtain predictions.

We use the standard formula $\hat{y} = X(X^T X)^{-1} X^T y$ with U_* instead of X and X instead of y:

$$U_* (U_*^T U_*)^{-1} U_*^T X = U_* (I)^{-1} U_*^T U \hat{D} V^T = U_* D_* V_*^T = \hat{X}$$

So! The $\hat{X}$ may be viewed as low-rank approximation of X from SVD or as matrix of forecasts of regressions of every original regressor x_j on r first principal components p_1, ..., p_r. Hence the sum of all squared approximation errors Q is the sum of all sum of squared residuals from these regressions

$$Q = \sum_{ij} (x_{ij} - \hat{x}_{ij})^2 = \|x_1 - \hat{x}_1\|^2 + \cdots + \|x_k - \hat{x}_k\|^2 = \text{SS}_1^{\text{res}} + \cdots + \text{SS}_k^{\text{res}}.$$

Let's calculate the R^2 in the first regression:

$$R_1^2 = 1 - \frac{\text{SS}_1^{\text{res}}}{\text{SST}_1} = 1 - \frac{\|x_1 - \hat{x}_1\|^2}{\|x_1\|^2} = 1 - \frac{\|x_1 - \hat{x}_1\|^2}{n-1}$$

The sum of all R_j^2 hence is

$$\sum R_j^2 = k - \frac{\sum (x_{ij} - \hat{x}_{ij})^2}{n-1}.$$

The average R_j^2, also called proportion of variance explained, is

$$\frac{\sum R_j^2}{k} = 1 - \frac{\sum (x_{ij} - \hat{x}_{ij})^2}{k(n-1)} = 1 - \frac{Q}{k(n-1)} = \frac{\sum_{i=1}^{r} d_{ii}^2}{\sum_{i=1}^{k} d_{ii}^2}.$$

How can we extract all these R_j^2 for different number of principal components? For simplicity let's consider x_1 and the first two principal components. The coefficient R_1^2 comes from regression of x_1 onto first $r = 2$ principal components p_1, p_2.

$$\hat{x}_1 = v_{11}p_1 + v_{12}p_2.$$

Artificial regressors p_1 and p_2 are orthogonal, hence R_1^2 may be decomposed into

$$R_1^2 = R_{11}^2 + R_{12}^2,$$

where R_{11}^2 comes from the regression of x_1 onto p_1 and R_{22}^2 — from the regression of x_1 onto p_2. The orthogonality of regressors also means that coefficient estimate v_{11} may be obtained from a univariate regression of x_1 onto p_1 alone. Similarly the regression coefficient v_{12} may be also obtained from a univariate regression of x_1 onto p_2 alone.

In a simple regression of original variable x_i onto principal component p_j the R_{ij}^2 may be obtained as a squared value of sample correlation between x_i and p_j. The sample correlation matrix of all columns X with all columns of P is called loadings matrix, L. The sample correlation is matrix is unaffected by scaling, so L is also a sample correlation matrix of columns of X and columns of U. Sample correlation matrix is just a sample covariance of standardized vectors. The columns in X are already standardized so we need to standardize only columns of P, or columns of U.

$$L = \begin{pmatrix} \text{s}\mathbb{C}\text{orr}(x_1, p_1) & \text{s}\mathbb{C}\text{orr}(x_1, p_2) & \dots \\ \text{s}\mathbb{C}\text{orr}(x_2, p_1) & \text{s}\mathbb{C}\text{orr}(x_2, p_2) & \dots \\ \vdots & \vdots & \vdots \end{pmatrix} = \frac{X^T P^{st}}{n-1} = \frac{X^T U^{st}}{n-1}$$

The meaning of its elements is threefold! First, ℓ_{ij} is the sample correlation between original x_i and principal component p_j, or normed principal component u_j or standardized principle component $\sqrt{n-1}u_j$. Second, ℓ_{ij} is the estimate of coefficient in the simple regression of x_i onto standardized principle component $\sqrt{n-1}u_j$. Third, as principal components are orthogonal, ℓ_{ij} is the estimate of coefficient in the multivariate regression of x_i onto all standardized components before the component $\sqrt{n-1}u_j$.

As $P^{st} = U^{st} = \sqrt{n-1}U$ we get

$$L = \frac{X^T U^{st}}{n-1} = \frac{VD^T U^T U\sqrt{n-1}}{n-1} = \frac{1}{\sqrt{n-1}} VD^T.$$

Hence, in the row 1 the sum of the first r squared values will give the coefficient R_1^2,

$$R_1^2 = \ell_{11}^2 + \ell_{12}^2 + \dots + \ell_{1r}^2.$$

If we continue summation over all k principal components we would obtain a perfect prediction of x_1,

$$\ell_{11}^2 + \ell_{12}^2 + \dots + \ell_{1k}^2 = 1.$$

Hence, $\|\text{row}_i\, L\| = 1$.

Another meaning of loadings may be seen from other view of SVD:

$$X = UDV^T = U(VD^T)^T = \sqrt{n-1}U(VD^T/\sqrt{n-1})^T = U^{st}L^T.$$

Hence, if we decompose original regressor x_i as a linear combination of standardized principal components then the weights will be given by the i-th row of the loadings matrix L.

The sample correlation matrix of columns of X is

$$\frac{X^T X}{n-1} = \frac{VD^T U^T UDV^T}{n-1} = \frac{VD^T DV^T}{n-1} = V\Lambda V^T.$$

Here Λ is the diagonalized sample correlation matrix of columns of X. On its diagonal we have $\lambda_i = d_{ii}^2/(n-1)$. The eigenvectors of the sample correlation matrix are just columns of V.

The loading matrix can be expressed in terms of V and Λ:

$$L = V\Lambda^{1/2} = V \cdot \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sqrt{\lambda_k} \end{pmatrix}.$$

As $\lambda_i = d_{ii}^2/(n-1)$ the average R_j^2 is also

$$\frac{\sum R_j^2}{k} = \frac{\sum_{i=1}^r d_{ii}^2}{\sum_{i=1}^k d_{ii}^2} = \frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^k \lambda_i}.$$

Two more facts about L matrix:

$$L^T L = \Lambda^{1/2} V^T V \Lambda^{1/2} = \Lambda^1 = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_k \end{pmatrix}$$

Hence $\ell_i \perp \ell_j$ for $i \neq j$, $\|\ell_j\|^2 = \lambda_j = d_{jj}/(n-1)$. The sample correlation matrix of columns of X may be written as

$$LL^T = \frac{1}{n-1} V D^T D V^T = \frac{X^T X}{n-1}$$

We again see that $(LL^T)_{ii} = 1$ as sample variance of standardized x_i.

Let's create a new variable q as a linear combination of columns of X with weight w, $q = Xw$.

We would like to maximize the sample variance of Xw while keeping $\|w\| = 1$. As sample variance for centered variable is just $\|Xw\|^2/(n-1)$ this is equivalent

$$\|Xw\|^2 \to \max_{\|w\|=1}$$

Solution 1 (no Lagrange, no matrix differential):

We can rewrite $\|Xw\|^2$ as

$$\|Xw\|^2 = (Xw)^T Xw = w^T X^T X w = w^T V D^T D V^T w$$

Let's introduce $a = V^T w$. Remark that $\|a\|^2 = a^T a = w^T V V^T w = w^T w = 1$ and

$$\|Xw\|^2 = a^T \cdot \begin{pmatrix} d_{11} & & \dots & \\ & d_{22} & \dots & \\ \dots & \dots & \dots & \dots \\ & & \dots & d_{kk} \end{pmatrix} \cdot a = d_{11}^2 a_1^2 + \dots + d_{kk} a_k^2$$

For $k = 3$ this may look like

$$\|Xw\|^2 = 1.3a_1^2 + 0.7a_2^2 + 0.2a_3^2.$$

As $\|a\|^2 = a_1^2 + \dots + a_k^2 = 1$ we conclude that the optimal point is $a = (1, 0, \dots, 0)$. And optimal w is $w = Va = v_1$, the first column of V.

Solution 2 (Rayleigh quotient):

$$\|Xw\|^2 \to \max_{\|w\|=1}$$

As $w^T w = 1$,

$$\frac{\|Xw\|^2}{n-1} = \frac{w^T V D^T D V^T w}{n-1} = \frac{w^T V D^T D V^T w}{(n-1)w^T w} = \frac{w^T \Lambda w}{w^T w}$$

Now take differential

$$d\frac{w^T \Lambda w}{w^T w} = \frac{\dots}{(w^T w)^2}$$

Solution 3 (with Lagrange)

Let's write the Lagrangian function wit Lagrange multiplier denoted μ:

$$\text{Lagrange}(w, \mu) = \|Xw\|^2 + \mu(1 - \|w\|^2).$$

The differential with respect to dw is

$$d\,\text{Lagrange} = 2w^T X^T X dw - 2\mu w^T dw.$$

First order conditions are

$$\begin{cases} 2w^T X^T X - 2\mu w^T = 0 \\ w^T w = 1 \end{cases}$$

Or, $X^T X w = \mu w$. Hence μ should be the eigenvalue of $X^T X$ matrix and w should be the eigenvector with $\|w\| = 1$. The optimal value of the function is

$$w^T X^T X w = w^T \mu w = \mu w^T w = \mu.$$

We should take the highest eigenvalue, that is $\mu_1 = (n-1)\lambda_1$ as λ_i are eigenvalues of the matrix $X^T X/(n-1)$.

Some useful sources:

https://www.math.chalmers.se/Stat/Grundutb/GU/MSA220/S19/extra/svd-relate
pdf

https://www.cs.columbia.edu/~djhsu/AML/lectures/notes-pca.pdf