Fundamentals of Image Processing

PCA - Detailed solution

Master 1 Computer Science - IMAge & VCC

Sorbonne Université

Year 2020-2021

1 Computation of the projected variance

- Data: $X_i \in \mathbb{R}^d$, $i = 1...n$, organized in a matrix X
- Average: $g = \frac{1}{n}\sum_{i=1}^{n} X_i$ (empirical estimation of the expectation) – A vector of dimension d.
- Projection operator: matrix $\Pi = vv^T$ where v is a unit vector (i.e. $v^Tv = 1$). The vector v defines a line on which data are projected. We want to find the best lines, maximizing the variance of the data along them.

Note that if X_i is decomposed as $X_i = \alpha v + \beta w$ where w defines the subspace orthogonal to v, then we get:

$$\Pi X_i = vv^T X_i = \alpha vv^T v + \beta vv^T w = \alpha v$$

since $v^T v = 1$ and $v^T w = 0$ since the subspace span by w is orthogonal to v.

Let us now compute the variance of the projection of the centered data (i.e. of the vectors $X_i - g$):

$$\begin{aligned}
\sigma_v^2(X) &= \frac{1}{n-1}\sum_{i=1}^{n}(vv^T(X_i - g))^T(vv^T(X_i - g)) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - g)^T vv^T vv^T (X_i - g) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - g)^T vv^T (X_i - g) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} v^T (X_i - g)(X_i - g)^T v \\
&= \frac{1}{n-1} v^T (\sum_{i=1}^{n}(X_i - g)(X_i - g)^T) v \\
&= v^T \Sigma v
\end{aligned}$$

Explanations:

- Line 1: definition of the variance, applied to $\Pi(X_i - g)$
- Line 2: use the fact that for any two matrices A and B, we have $(AB)^T = B^T A^T$
- Line 3: use $v^T v = 1$ (v is a unit vector)
- Line 4: $(X_i - g)^T v$ is the scalar product of the vectors $X_i - g$ and v, which is symmetrical, hence equal to the scalar product of v and $X_i - g$, which writes $v^T(X_i - g)$
- Line 5: uses the linearity and the fact that v does not depend on i
- Line 6: $\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - g)(X_i - g)^T$ which is the covariance matrix of the data

2 Derivation of a bilinear form

Let us prove that $\frac{\partial(v^T \Sigma v)}{\partial v} = 2\Sigma v$, where $\frac{\partial}{\partial v}$ denotes the vector of $\frac{\partial}{\partial v_k}, k = 1...d$.

Let $v = (v_1....v_d)^T$, s_{ij} the coefficients of matrix Σ (with $s_{ij} = s_{ji}$). We have:

$$v^T \Sigma v = \sum_{i=1}^{d} \sum_{j=1}^{d} s_{ij} v_i v_j$$

When taking the derivative with respect to v_k, only the terms involving v_k will be non zero. This leads to:

$$\frac{\partial(v^T \Sigma v)}{\partial v_k} = \sum_{i=1}^{d} s_{ik} v_i + \sum_{j=1}^{d} s_{kj} v_j = 2 \sum_{i=1}^{d} s_{ik} v_i = 2\Sigma_k v$$

where Σ_k denotes the k^{th} line of Σ. Hence $\frac{\partial(v^T \Sigma v)}{\partial v} = 2\Sigma v$.

3 Solving PCA

We want to find v maximizing $\sigma_v^2(X)$ under the constraint $v^T v = 1$. Using Lagrange multiplier, this leads to the maximization of

$$v^T \Sigma v + \lambda(1 - v^T v)$$

Setting that the derivative of this functional is equal to 0 leads to:

$$2\Sigma v - 2\lambda v = 0$$

i.e.

$$\Sigma v = \lambda v$$

This means that v is an eigenvector of Σ, associated with the eigenvalue value λ. Now we have

$$\sigma_v^2(X) = \lambda v^T v = \lambda$$

which means that the projected variance on v is exactly the eigenvalue associated with v. Maximizing this variance amounts to choose the largest eigenvalue. In practice PCA reduces the dimension of the feature space by choosing the q highest eigenvalues and projecting the data on the subspace span by the corresponding q eigenvectors.

Note that Σ is a real symmetric matrix, which is definite positive. It is therefore diagonalizable in an orthonormal basis (the basis of eigenvectors), and the eigenvalues are positive.