

1 Derivation of PCA

Assume we are given n training data points (\mathbf{x}_i, y_i) . We collect the target values into $\mathbf{y} \in \mathbb{R}^n$, and the inputs into the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where the rows are the d -dimensional feature vectors \mathbf{x}_i^\top corresponding to each training point. Furthermore, assume that the data has been centered such that $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$, $n > d$ and \mathbf{X} has rank d . The covariance matrix is given by

$$\Sigma = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

When $\bar{\mathbf{x}} = \mathbf{0}$ (i.e., we have subtracted the mean in our samples), we obtain $\Sigma = \frac{1}{n} \mathbf{X}^\top \mathbf{X}$.

- (a) Maximum Projected Variance: We would like the vector \mathbf{w} such that projecting your data onto \mathbf{w} will retain the maximum amount of information, i.e., variance. We can formulate the optimization problem as

$$\max_{\mathbf{w}: \|\mathbf{w}\|_2=1} \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{w})^2 = \max_{\mathbf{w}: \|\mathbf{w}\|_2=1} \frac{1}{n} \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}. \quad (1)$$

Show that the maximizer for this problem is equal to the eigenvector \mathbf{v}_1 that corresponds to the largest eigenvalue λ_1 of Σ . Also show that optimal value of this problem is equal to λ_1 .

Hint: Use the spectral decomposition of Σ and consider reformulating the optimization problem using a new variable.

- (b) Let us call the solution of the above part \mathbf{w}_1 . Next, we will use a *greedy procedure* to find the i th component of PCA by doing the following optimization

$$\begin{aligned} & \text{maximize} && \mathbf{w}_i^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}_i \\ & \text{subject to} && \mathbf{w}_i^\top \mathbf{w}_i = 1 \\ & && \mathbf{w}_i^\top \mathbf{w}_j = 0 \quad \forall j < i, \end{aligned} \tag{2}$$

where $\mathbf{w}_j, j < i$ are defined recursively using the same maximization procedure above. Show that the maximizer for this problem is equal to the eigenvector \mathbf{v}_i that corresponds to the i th eigenvalue λ_i of Σ . Also show that optimal value of this problem is equal to λ_i .

2 Ridge regression vs. PCA

In this problem we want to compare two procedures: The first is ridge regression with hyperparameter λ , while the second is applying ordinary least squares after using PCA to reduce the feature dimension from d to k (we give this latter approach the short-hand name k -PCA-OLS where k is the hyperparameter).

Notation: The singular value decomposition of \mathbf{X} reads $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{n \times n}$, $\mathbf{\Sigma} \in \mathbb{R}^{n \times d}$ and $\mathbf{V} \in \mathbb{R}^{d \times d}$. We denote by \mathbf{u}_i the n -dimensional column vectors of \mathbf{U} and by \mathbf{v}_i the d -dimensional column vectors of \mathbf{V} . Furthermore the diagonal entries $\sigma_i = \Sigma_{i,i}$ of $\mathbf{\Sigma}$ satisfy $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_d > 0$. For notational convenience, assume that $\sigma_i = 0$ for $i > d$.

(a) Consider running ridge regression with $\lambda > 0$ in the \mathbf{V} -transformed coordinates, i.e.,

$$\widehat{\mathbf{w}}_{\text{ridge}} = \arg \min_{\mathbf{w}} \|\mathbf{X}\mathbf{V}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2.$$

Note that this does not correspond to any dimensionality reduction, just a change of variables. It turns out that the solution in this case can be written as:

$$\widehat{\mathbf{w}}_{\text{ridge}} = \text{diag}\left(\frac{\sigma_1}{\lambda + \sigma_1^2}, \dots, \frac{\sigma_d}{\lambda + \sigma_d^2}, 0, \dots, 0\right) \mathbf{U}^\top \mathbf{y}. \quad (3)$$

Use $\widehat{y}_{\text{test}} = \mathbf{x}_{\text{test}}^\top \mathbf{V} \widehat{\mathbf{w}}_{\text{ridge}}$ to denote the resulting prediction for a hypothetical \mathbf{x}_{test} . Using (3) and the appropriate scalar $\{\beta_i\}$, show that this prediction can be written as:

$$\widehat{y}_{\text{test}} = \mathbf{x}_{\text{test}}^\top \sum_{i=1}^d \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \quad (4)$$

(b) Suppose that we do k -PCA-OLS — i.e. ordinary least squares on the reduced k -dimensional feature space obtained by projecting the raw feature vectors onto the $k < d$ principal components of $\mathbf{\Sigma}$. Use $\widehat{y}_{\text{test}}$ to denote the resulting prediction for a hypothetical \mathbf{x}_{test} .

It turns out that the learned k-PCA-OLS predictor can also be written as:

$$\widehat{y}_{test} = \mathbf{x}_{test}^\top \sum_{i=1}^d \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \quad (5)$$

What are the $\beta_i \in \mathbb{R}$ coefficients in this case?

Hint: Some of these β_i will be zero.

(c) Compare $\widehat{\mathbf{y}}_{\text{PCA}}$ with $\widehat{\mathbf{y}}_{\text{ridge}}$ (at different λ), how do you find their relationship?