

This discussion was released **Friday, October 9**.

This discussion material is on PCA, LASSO, ridge regression and their relationships. We will first start with this [Jupyter notebook](#). We will then come back to the document and consider the relationship between PCA and ridge regression. Note that this is also part of your homework problem.

As a reminder, if you have questions, we will answer them via the queue at oh.eecs189.org. Once you complete the Jupyter notebook, please return to this worksheet.

1 Jupyter Notebook

2 Ridge regression vs. 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 $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$, $n > d$ and \mathbf{X} has rank d .

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 $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_d \geq 0$. For notational convenience, assume that $\sigma_i = 0$ for $i > d$.

(a) It turns out that the ridge regression optimizer (with $\lambda > 0$) in the \mathbf{V} -transformed coordinates

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

has the following expression:

$$\widehat{\mathbf{w}}_{\text{ridge}} = \text{diag}\left(\frac{\sigma_i}{\lambda + \sigma_i^2}\right) \mathbf{U}^\top \mathbf{y}. \quad (1)$$

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 (1) and the appropriate scalar $\{\beta_i\}$, this can be written as:

$$\widehat{y}_{\text{test}} = \mathbf{x}_{\text{test}}^\top \sum_{i=1}^d \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \quad (2)$$

What are the $\beta_i \in \mathbb{R}$ for this to correspond to (1) from ridge regression?

- (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 the covariance matrix $\mathbf{X}^\top \mathbf{X}$. Use $\widehat{\mathbf{y}}_{test}$ to denote the resulting prediction for a hypothetical \mathbf{x}_{test} ,

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

$$\widehat{\mathbf{y}}_{test} = \mathbf{x}_{test}^\top \sum_{i=1}^d \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \quad (3)$$

Give the $\beta_i \in \mathbb{R}$ coefficients for k-PCA-OLS. Show work.

Hint 1: some of these β_i will be zero. Also, if you want to use the compact form of the SVD, feel free to do so if that speeds up your derivation.

Hint 2: some inspiration may be possible by looking at the next part for an implicit clue as to what the answer might be.

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

Contributors:

- Anant Sahai
- Fanny Yang
- Peter Wang
- Philipp Moritz